

Progress monitoring and data-based decision-making in inclusive schools

Edited by

Markus Gebhardt, Stefan Blumenthal, David Scheer, Yvonne Blumenthal, Sarah Powell and Erica Lembke

Published in

Frontiers in Education
Frontiers in Psychology



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-2378-0
DOI 10.3389/978-2-8325-2378-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Progress monitoring and data-based decision-making in inclusive schools

Topic editors

Markus Gebhardt — University of Regensburg, Germany

Stefan Blumenthal — University of Rostock, Germany

David Scheer — Ludwigsburg University of Education, Germany

Yvonne Blumenthal — University of Rostock, Germany

Sarah Powell — The University of Texas at Austin, United States

Erica Lembke — University of Missouri, United States

Citation

Gebhardt, M., Blumenthal, S., Scheer, D., Blumenthal, Y., Powell, S., Lembke, E., eds. (2023). *Progress monitoring and data-based decision-making in inclusive schools*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2378-0

Table of contents

- 05 **Editorial: Progress monitoring and data-based decision-making in inclusive schools**
Markus Gebhardt, Stefan Blumenthal, David Scheer, Yvonne Blumenthal, Sarah Powell and Erica Lembke
- 07 **Generalizability of Written Expression Curriculum-Based-Measurement in the German Language: What Are the Major Sources of Variability?**
Julia Winkes and Pascale Schaller
- 20 **Statistical Power of Piecewise Regression Analyses of Single-Case Experimental Studies Addressing Behavior Problems**
Jürgen Wilbert, Moritz Börnert-Ringleb and Timo Lüke
- 33 **Individual, generalized, and moderated effects of the good behavior game on at-risk primary school students: A multilevel multiple baseline study using behavioral progress monitoring**
Tatjana Leidig, Gino Casale, Jürgen Wilbert, Thomas Hennemann, Robert J. Volpe, Amy Briesch and Michael Grosche
- 48 **Monitoring indicators of scholarly language: A progress monitoring tool for documenting changes in narrative complexity over time**
Megan Israelsen-Augenstein, Carly Fox, Sandra L. Gillam, Sarai Holbrook and Ronald Gillam
- 62 **To use or not to use learning data: A survey study to explain German primary school teachers' usage of data from digital learning platforms for purposes of individualization**
Alina Hase, Leonie Kahnbach, Poldi Kuhl and Dirk Lehr
- 77 **Understanding and improving teachers' graph literacy for data-based decision-making *via* video intervention**
Jana Jungjohann, Markus Gebhardt and David Scheer
- 95 **Developing an online system to support algebra progress monitoring: Teacher use and feedback**
Pamela M. Stecker and Anne Foegen
- 115 **Developing progress monitoring measures: Parallel test construction from the item-up**
Leanne R. Ketterlin-Geller, Anthony Sparks and Jennifer McMurrer
- 129 **Teachers' visual inspection of Curriculum-Based Measurement progress graphs: An exploratory, descriptive eye-tracking study**
Roxette M. van den Bosch, Christine A. Espin, Maria T. Sikkema-de Jong, Siuman Chung, Priscilla D. M. Boender and Nadira Saab

- 142 **Multilevel and empirical reliability estimates of learning growth: A simulation study and empirical illustration**
Boris Forthmann, Natalie Förster and Elmar Souvignier
- 154 **Students' learning growth in mental addition and subtraction: Results from a learning progress monitoring approach**
Sven Anderson, Michael Schurig, Daniel Sommerhoff and Markus Gebhardt
- 172 **Continuous norming in learning progress monitoring—An example for a test in spelling from grade 2–4**
Michael Schurig, Stefan Blumenthal and Markus Gebhardt
- 190 **The BEHAVE application as a tool to monitor inclusive interventions for subjects with neurodevelopmental disorders**
Gianluca Merlo, Antonella Chifari, Giuseppe Chiazzese, Paola Denaro, Noemi Firrer, Nicola Lo Savio, Simona Patti, Luisa Palmegiano, Davide Taibi and Luciano Seta



OPEN ACCESS

EDITED AND REVIEWED BY
Douglas F. Kauffman,
Medical University of the Americas–Nevis,
United States

*CORRESPONDENCE
Markus Gebhardt
✉ markus.gebhardt@ur.de

RECEIVED 14 March 2023
ACCEPTED 13 April 2023
PUBLISHED 25 April 2023

CITATION
Gebhardt M, Blumenthal S, Scheer D,
Blumenthal Y, Powell S and Lembke E (2023)
Editorial: Progress monitoring and data-based
decision-making in inclusive schools.
Front. Educ. 8:1186326.
doi: 10.3389/feduc.2023.1186326

COPYRIGHT
© 2023 Gebhardt, Blumenthal, Scheer,
Blumenthal, Powell and Lembke. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Editorial: Progress monitoring and data-based decision-making in inclusive schools

Markus Gebhardt^{1*}, Stefan Blumenthal², David Scheer³,
Yvonne Blumenthal², Sarah Powell⁴ and Erica Lembke⁵

¹Faculty of Human Sciences, University of Regensburg, Regensburg, Germany, ²Faculty of Philosophy, University of Rostock, Rostock, Mecklenburg-Vorpommern, Germany, ³Faculty of Special Education, Ludwigsburg University of Education, Ludwigsburg, Baden-Württemberg, Germany, ⁴Department of Special Education, The University of Texas at Austin, Austin, TX, United States, ⁵Department of Special Education, University of Missouri, Columbia, KY, United States

KEYWORDS

CBM, progress monitoring (PM) measures, learning growth, teacher feedback, progress graph, single case analysis, competence and performance

Editorial on the Research Topic

Progress monitoring and data-based decision-making in inclusive schools

Despite extensive research and positive practices related to inclusive education, some students still struggle with academic skills. Progress monitoring (PM) is a valuable approach that can provide explicit feedback to teachers in schools about how students respond to instruction. The fundamental idea behind PM is to document the learning development of students and use the data to inform instructional decisions about interventions over time, using repeated, brief, and reliable standardized tests. PM is a formative diagnostic that enables the measurement and evaluation of learning development at multiple points in time, providing feedback to teachers and learners. Unlike summative assessment, which evaluates learning outcomes, PM aims to measure for the purpose of supporting learning.

Without PM, students with exceptional needs may be evaluated based on their performance relative to their classmates, rather than their own individual progress. Research indicates that when teachers use PM, positive effects on student outcomes can be seen. However, the use of PM is not widespread, which may be due to teachers having additional work or a lack of knowledge on how to use PM in the context of data-based decision making. Additionally, tests or online platforms for PM are not available in many countries and languages.

To effectively measure learning progress, PM measures must provide both the psychometric quality criteria for status tests and the quality to measure learning progress. Classical test theory is no longer sufficient for this purpose since learning trajectories differ among students. PM measures must be uniform over time, both for an individual student and for specific groups of students (measurement invariance), and PM must be sensitive to learning trajectories (i.e., sensitive to change, even for weak learners). Moreover, PM measures must be brief and easy to use so they can be used frequently in everyday teaching. Therefore, it is crucial that PM measures are practical, useful, and economical. This is because PM can only be effective when teachers reflect on their instructional decisions based on the new information provided by the PM data. Compared to status tests, the requirements of PM are much higher, both psychometrically and in terms of practical implementation.

Therefore, PM should be supported by adapted materials and recommendations to aid teachers and students.

In the field of education, a range of studies have been conducted to improve the reliability and effectiveness of various methods used for assessment, monitoring and evaluation of student progress. The following studies are a few examples of such efforts.

Methods

Wilbert et al. conducted a study to analyze the statistical power of piecewise regression analyses in single-case experimental studies. Their research demonstrated that this method can be a useful tool for planning and assessing single case studies, which are crucial for reviewing evidence-based practice.

Forthmann et al. conducted a simulation study to assess the reliability of measures used for monitoring student progress. They found that reliability estimation works well across a variety of simulation conditions, but it can be biased under certain circumstances, such as when data quality is very poor or empirical reliability is estimated.

Ketterlin-Geller et al. described an approach to adapting Automated Item Generation (AIG) principles to develop parallel progress monitoring measures.

Schurig et al. presented a study on continuous norming in learning progress monitoring for a spelling test. Their data was obtained through a longitudinal study of students in grades 2 to 4.

Test construction

Anderson et al. conducted a longitudinal study on mental computation over a period of 34 weeks with data collected for 12 measurement intervals. Their research was affected by the COVID-19 pandemic.

Israelsen-Augenstein et al. developed a new measure, the Monitoring Indicators of Scholarly Language (MISL), which was shown to be a valid measure of narrative production abilities.

Winkes and Schaller developed a written expression curriculum-based measurement (CBM-W) suitable as a universal screening tool but not for progress monitoring of individual students.

Case studies

Leidig et al. conducted a study on the impact of the Good Behavior Game (GBG) on at-risk students' academic engagement

and disruptive behavior. They used behavioral progress monitoring with a multiple baseline design in a German inclusive primary school sample.

Merlo et al. introduced a tool called BEHAVE to monitor inclusive interventions and presented two case studies involving kindergarten children with neurodevelopmental disorders.

Teacher training

Stecker and Foegen developed an online system to support algebra progress monitoring and determined that it improved teachers' scoring in algebra measures based on online instruction.

Jungjohann et al. developed a video intervention for linear trend identification using Tukey Tri-Split and demonstrated that the video instruction is more effective than text-based hints.

Hase et al. conducted an online survey study on the usage of learning data from Digital Learning Platforms (DLP).

Van den Bosch et al. examined teachers' visual inspection of Curriculum-Based Measurement (CBM) progress graphs using eye-tracking technology. Their study revealed variability in teachers' patterns of graph inspection, which was linked to their abilities to describe the graphs.

Author contributions

MG set up a draft and all authors improved it. All authors participated equally in writing the editorial. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



Generalizability of Written Expression Curriculum-Based-Measurement in the German Language: What Are the Major Sources of Variability?

Julia Winkes^{1*†} and Pascale Schaller^{2†}

¹ Department of Special Education, University of Fribourg, Fribourg, Switzerland, ² Institute of Primary Education, University of Teacher Education Bern, Bern, Switzerland

OPEN ACCESS

Edited by:

Markus Gebhardt,
University of Regensburg, Germany

Reviewed by:

Sterett Mercer,
The University of British Columbia,
Canada
Stefanie Roos,
University of Siegen, Germany

*Correspondence:

Julia Winkes
julia.winkes@unifr.ch

†ORCID:

Julia Winkes
orcid.org/0000-0002-4383-1926
Pascale Schaller
orcid.org/0000-0002-3600-5386

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

Received: 13 April 2022

Accepted: 23 May 2022

Published: 15 June 2022

Citation:

Winkes J and Schaller P (2022)
Generalizability of Written Expression
Curriculum-Based-Measurement
in the German Language: What Are
the Major Sources of Variability?
Front. Educ. 7:919756.
doi: 10.3389/feduc.2022.919756

This study aimed to identify the sources of measurement error that contribute to the intraindividual variability of written expression curriculum-based measurement (CBM-W) and assess how many German writing samples of 3 or 5 min duration are necessary to make sufficiently reliable relative and absolute decisions. Students in grade 3 ($N = 128$) and grade 6 ($N = 118$) wrote five CBM-W probes of 5 min each within 1 week, which were scored for commonly used metrics (i.e., words written, correct writing sequences). Analyses within the generalizability theory framework showed that between-student differences accounted for 36–60% of the variance. The student \times writing prompt interaction was the largest source of variability, particularly among younger students (44%), while writing prompt *per se* and writing time explained no variance. Two to four writing samples of 3 min are sufficient for most scoring methods to achieve relative reliability >0.80 . CBM-W in German proved inadequate for the grade levels studied for absolute decisions. These findings imply that CBM-W in this form in German-speaking primary grades is suitable as a universal screening tool but not as a tool for progress monitoring of individual students.

Keywords: curriculum-based measurement (CBM), writing, generalizability theory, reliability, variability

INTRODUCTION

Although writing is a crucial competence for students' academic and professional success (Traga Philippakos and FitzPatrick, 2018, p. 165), The National Commission on Writing in America's Schools and Colleges (2003) designated it a neglected basic skill. This wake-up call was a response to the National Assessments of Educational Progress, which captured many students who did not reach a proficient writing level. In 2011, for example, 74% of eighth-graders scored at the "basic" or "below basic" levels, and only 3% could be described as advanced writers compared to their grade level requirements (National Center for Education Statistics, 2011). So, in addition to students with a learning disability, there are a significant number of low achieving writers who lack writing proficiency (Graham and Perin, 2007). Until now, writing at the text level has hardly been included in national assessments in German speaking countries, with the exception of the DESI study (DESI-Konsortium, 2006). It showed that at grade 9, about 29% of the students are not able to formulate a letter adequately for the addressee and that the linguistic quality of these students' writing is also extremely low. Thus, although the educational system, curriculum, teaching methods, and orthography to be learned differ between German-speaking and English-speaking countries, it can be surmised that, as in English-speaking countries, weak writing skills are present

but probably underdiagnosed in German-speaking countries. The problem is exacerbated by the lack of standardized writing assessments in German, so that writing is usually only systematically evaluated at the spelling level and cannot be reliably assessed at the text level. Struggling writers produce texts that are generally shorter, less interesting, and poorly organized at the sentence and paragraph level (Hooper et al., 2002). The children's texts are marred by inordinate numbers of mechanical, spelling, and grammatical errors (Dockrell et al., 2015). Therefore, the difficulties of these children go far beyond pure spelling problems since the spelling is only a small part of the skills required to produce linguistically correct and content-appropriate texts of good quality. A competency that is an indicator of writing quality at the text level is writing fluency (Kim et al., 2017; Poch et al., 2021). At the same time, writing fluency proves to be sensitive to change since both speed/productivity and accuracy increase with a growing writing routine. Skills that serve as indicators of general performance in an academic area are useful as vital signs for screening students at risk and for progress monitoring (Fuchs, 2004, 2017). For this purpose, short, reliable, and valid learning samples are used in curriculum-based measurements (CBM), which capture critical skills simply and economically (Deno, 1985).

CBM Writing (CBM-W), as an indicator of writing proficiency, uses short writing samples for this aim. The students are given writing prompts, such as pictures or introductory sentences and asked to write for 3 or 5 min. Various scoring methods are available, such as the number of words written (TWW), the number of words spelled correctly (WSC), the number of correct writing sequences (CWS), or the number of correct minus incorrect writing sequences (CIWS). Thus, the collected measures do not focus on content-related text quality (e.g., ideation or genre specificity) but either on writing quantity (TWW), spelling (WSC), or linguistic units whose correct realization requires the integration of individual sub-competencies (writing motor skills, retrieval of linguistic knowledge, semantics and spelling) (CWS and CIWS).

Since the beginning of CBM research, great importance has been attached to ensuring that the methods used reflect the learners' performance in reliable ways – despite their easy handling and the short time for implementation and scoring (Fuchs et al., 1983). Reliable information is key because it builds the foundation for the teachers' important (high stakes and low stakes) data-based decisions (McMaster and Espin, 2007). Parallel forms are needed in their functions as repeated screenings and progress monitoring. These require high parallel test reliability (rank-ordering of students) and stability (consistent within-student performance over time) (Campbell et al., 2013). This central claim contrasts with an observation we made in a previous research project on CBM-W (Winkes and Schaller, 2022). In this study, students in grades 3–6 wrote ten writing samples within a short period of 2 weeks. Parallel test reliability was satisfactory overall, but a closer look at the children's test data revealed considerable intraindividual variability between student test scores. We found this observation remarkable because the CBM samples were collected within a quite short period. In general, meaningful variation in performance within individuals

is not fundamentally new for CBM (Christ et al., 2016). It invites a closer look at the issue of “variability” – here specific to CBM-W. Accordingly, the present study aims to understand the sources of this variability in more detail and examine the influence of story starter, rater/class, and length of writing sample on the generalizability of CBM-W in German.

Potential Sources of Variability in Written Expression Curriculum-Based Measurement

Taking the object of learning as a starting point, increased intraindividual variability in writing, compared to other performance areas such as reading, spelling, or mathematics, is not necessarily surprising. On the one hand, variability can be understood as an expression of the complexity of the writing process itself. Text writing is a problem-solving process that requires the integration of different hierarchy-low and hierarchy-high processing skills and thus does not succeed with equal fluency and quality at all times (Alamargot and Chanquoy, 2001; Kent and Wanzek, 2016). On the other hand, writing a text is a creative language-productive task, which leads to a special starting point. In other areas of CBM, the number of given items (e.g., arithmetic problems, words to be read) that can be correctly solved in a defined time is usually recorded. In writing, on the other hand, the items to be assessed are produced by the child himself.

Two children with the same writing skills will arrive at two very different final products based on the same story starter. The same is true when testing a child repeatedly. Even using the same story starter and under comparable contextual conditions, a child is unlikely to use the same words and phrases to write a story on two different occasions. As Ritchey et al. (2016) point out, writing opens up opportunities for students to actively avoid difficult words or choose simpler words and sentence structures, which influences the difficulty of different texts.

A certain variability is, therefore, to be expected, which is inherent to the writing process itself and which is caused by the open nature of the task. For this reason, it is particularly important in writing to design the conditions for progress-monitoring measurements so that as many external sources of measurement error as possible can be reduced and that as much of the remaining variance as possible can be attributed to the subject itself. Potential sources of measurement error could include, for example, the different story starters, the length of the writing time, or the rater. In the following, we discuss the state of knowledge regarding the importance of these factors concerning CBM-W.

The Task or Writing Prompt

So far, the role of writing prompts has been surprisingly little investigated in CBM-W. Existing studies on this topic focus on what kinds of writing prompts are appropriate at which grade level. For example, various word- and sentence-level task formats have been suggested for beginning writers, requiring text production in response to a picture or story starter with descriptive or narrative content (Ritchey et al., 2016). In

the higher grades, the question arises in particular whether expository or narrative prompts better represent students' academic writing abilities, as they are potentially more in line with typical school writing tasks [for two recent meta-analyses related to the validity of different writing genres, see Romig et al. (2017, 2020)]. Within the different genres (e.g., expository vs. narrative), it is assumed that different tasks are comparable and that the writing prompts used are equivalent, without this assumption having been sufficiently tested empirically to date (Keller-Margulis et al., 2016a). In contrast, for other forms of CBM (e.g., reading fluency; mathematics), great importance is attached to the development of parallel test versions. As Christ et al. (2016) describe, the variability of student performance across forms in CBM research has led to the standardization not only of the procedures for administration and scoring but also of the materials used. This development does not seem to have established itself specifically in writing. While collections of tasks are available at CBM-W^{1, 2}, it is also possible for practitioners to invent story starters themselves, as long as they are age-appropriate and do not evoke a one-word response (Hosp et al., 2016). However, McMaster and Espin (2007) point out that students' background knowledge and interest in different writing prompts may vary greatly, affecting the quality and quantity of their writing. Existing studies of writing prompt comparability use alternate-form reliability to examine how closely different writing samples correlate with each other [see for grades 1–5 the studies of Gansle et al. (2002), Weissenburger and Espin (2005), Gansle et al. (2006), Campbell et al. (2013), and Allen et al. (2019)]. They usually set a Pearson's correlation coefficient of $r \geq 0.70$ for sufficient reliability in CBM-W (Allen et al., 2019, p. 10). The various scores usually reach this threshold. However, Allen et al. (2019) found large differences between the correlation coefficients. For example, for grade level 3, the CIWS coefficients vary between 0.31 and 0.92, and for TWW, between 0.50 and 0.91. McMaster and Espin (2007, p. 69) point out that the standards for reliability coefficients should possibly be set domain-dependently. For CBM of oral reading fluency, reliability coefficients of $r > 0.85$ are usually reported. Such high coefficients are not expected for CBM-W, which is probably related to the test setting: A text as a continuation of a story starter can take an infinite number of possible forms, which is not the case for reading fluency. Moreover, the procedure established in CBM-W for eliciting parallel test reliability, namely calculating the correlations between several CBM tests administered simultaneously, only verifies part of the necessary conditions for parallel tests. These should also have equal means and variances (Christ and Hintze, 2007). This assumption has not yet been controlled for CBM-W.

The Writing Time

The main characteristic of progress monitoring and CBM procedures is that they are highly time-efficient in implementation and evaluation (Deno, 2003). This is the only way to ensure that regular use is possible in everyday school life,

especially if used in parallel in several performance areas (e.g., reading, spelling, writing, mathematics). Thus, the duration of CBM-W should be as short as possible but as long as necessary to ensure a sufficiently reliable capture of the feature to be examined. Most studies on CBM-W refer to 3-min writing samples preceded by a planning period of 1 min, and this procedure is also the standard in practice (Hosp and Kaldenberg, 2020). However, the effects of increased writing time (e.g., 5, 7, or 10 min) on the reliability of measures in CBM-W have been studied on several occasions. Younger students showed only slight differences in the reliability of shorter and longer writing samples (Espin et al., 2000). For older students, increasing the writing time to 5–7 min was necessary to achieve reliability > 0.70 (Weissenburger and Espin, 2005; Campbell et al., 2013), which was also true for the English language learners (ELL; Espin et al., 2008). It is still unclear up to which grade level a writing time of 3 min is sufficient and from when the writing time should be increased. Of course, the choice of writing duration also depends on the purpose. Espin et al. (2008) recommend a 7-min writing sample for older students due to increased reliability if CBM-W is used as a screening only one to three times per school year. For use at shorter and more regular intervals (e.g., once per week), they recommend a more economical 5-min writing sample.

The Rater

Since CBM-W evaluates texts using different scores, the question arises as to what role the rater's influence plays in the results. Campbell et al. (2013) report very high interrater reliabilities: they indicate average interscorer agreement from 80% (CIWS) to 99% (TWW). The differences between scores that report text volume (TWW) and scores that address writing accuracy can plausibly be explained because, in TWW, only the words are counted, whereas CWS or CIWS assess the correctness of writing sequences. Different ratings of the same writing sequence are sometimes related to the fact that different raters assume different target structures announced by the child. The very high interrater reliabilities also for CWS and CIWS [see, Weissenburger and Espin (2005), Gansle et al. (2006), Campbell et al. (2013), and Keller-Margulis et al. (2016b)] are probably due to intensive training of raters, which cannot be assumed in the practical application of CBM-W.

Generalizability Theory

Studies on the psychometric properties of CBM-W have so far almost exclusively used the framework of Classical Test Theory (CTT) by investigating parameters such as parallel test reliability, interrater reliability, or criterion validity. Especially in the context of progress monitoring, where an idiographic reference norm is usually used, Generalizability Theory (G-theory) provides an alternative. It has three advantages: First, it can investigate different sources of measurement error simultaneously. It uses repeated measures ANOVA to estimate the variance components for each source of variation (referred to as facets in G-theory terminology) in the observed values and the interactions among these facets. Thus, G-theory provides a good overview of the main contributors to measurement error, which, unlike in CTT, are analyzed in the same model. This information

¹ www.aimsweb.com

² www.interventioncentral.org

can subsequently be used to effectively optimize assessment procedures (Hintze et al., 2000).

The second advantage of G-theory concerns the reliability coefficients reported. In CTT, the calculation of parallel test reliability examines whether a child moves in the same rank relative to the other children in the group on repeated performance measures within the subject group, such as the class. If, for example, the weakest child in the class always achieves the lowest measurement result in the class over five measurement points, then classical test theory evaluates this as an indication of high parallel test reliability, although the child's competence values may vary greatly between these five measurement points (see Keller-Margulis et al., 2016a). In G-theory, there is a corresponding coefficient of generalizability (G-coefficient) to this classical reliability coefficient, which is thus informative for relative decisions related to the ranking of subjects (Cardinet, 1998).

In addition, the dependability-coefficient (D-Coefficient) is another parameter that focuses on the performance level, independent of the ranking. It can be used to make absolute, criterion-referenced decisions. D-coefficients are more conservative than G-coefficients for this reason. They are particularly suitable for use in progress monitoring, as Fan and Hansmann (2015) argue: "... research has acknowledged that having high-rank order reliability at a group design level (like the generalizability coefficient in G theory) cannot guarantee the comparability of CBM-R scores used at the individual student level" (S. 207). The minimum thresholds of G- and D-coefficients depend on the application situation. For low-stakes decisions, a reliability of 0.80 is considered sufficient and feasible in practice. However, for high-stakes decisions it is usually argued referring to Nunnally (1967) that coefficients below 0.90 are unacceptable (Graham et al., 2016; Keller-Margulis et al., 2016a; Kim et al., 2017; Wilson et al., 2019). The third advantage of G-theory is that G-coefficients and D-coefficients can not only be generated for the actual conditions of investigation but it can be estimated with the help of so-called decision studies (D-studies) how these coefficients vary under other conditions. This allows identifying the minimum requirements to obtain sufficiently high measurement reliability. For example, how many writing samples of what duration are necessary to achieve reliability above 0.80 for relative or absolute decisions can be checked.

Use of Generalizability Theory in Written Expression Curriculum-Based Measurement

The advantages of G-theory over CTT lead to popularity in writing assessments. A recent review of the content of the journal "Assessing Writing" from 2000 to 2018 (Zheng and Yu, 2019) indicates that G-theory was the most frequently used method during this period. However, existing studies mainly examined college students or adult L2 learners. Which factors influence the reliability of writing scores in children has not yet been much addressed (Kim et al., 2017). Specifically, for CBM-W, generalizability theory has been used only twice: In the study of Keller-Margulis et al. (2016a), 2nd–5th grade students

wrote three 7-min writing samples at three time points each year. After each minute, subjects changed the color of their pen while writing so that the impact of writing time on the reliability of the measures could be assessed (from 1 to 7 min). Other facets included students (between student differences), story starter, benchmark (time within a year), and interactions among these factors. Nearly half of the variance in CBM-W proved to be the non-systematic error. Reliability above 0.80 – as the threshold for low-stakes decisions – was achieved with the relative reliability coefficient at most grade levels by three 3-min writing samples, the D-coefficient for absolute decisions reached the threshold of 0.80 with two 5-min or three 4-min tests. For contexts with high stakes decisions, depending on grade level and scoring method, three 5- to 7-min writing samples were needed for sufficient relative reliability above 0.90, and three 7-min writing samples were necessary for sufficient absolute reliability. Thus, the typical CBM-W implementation convention of using a single writing sample of 3 min as a screening instrument proves inadequate. The use of multiple longer writing samples, on the other hand, severely limits the feasibility of CBM-W in its function as a screening, making widespread implementation unrealistic for many schools. Therefore, the authors are skeptical about whether CBM-W is the best way to identify at-risk students in writing.

In the second study, which used G-theory, Kim et al. (2017) examined the influence of rater ($N = 2$) and task ($N = 3$) on the reliability of writing tasks in expository and narrative genres for 3rd and 4th-grade students. The writing time here was 15 min per text, so the task does not correspond to conventional implementation conditions for CBM-W, but the texts were analyzed using the scoring methods for CBM-W, among others. For the evaluation *via* TWW and CWS, it was found that most of the variance was explained by the person (57–69%) and another large proportion by the interaction between person and task (31–41%). Variability was minimal when explained by rater, person \times rater, or the non-systematic error. Subsequent D-studies indicated that for both absolute and relative decisions, two to four tasks and a single rater were necessary to reach the criterion of 0.80 and five to six tasks and one rater were necessary for the criterion of 0.90 reliability.

The Present Study

The present study explores the major sources of variability of CBM-W in German in grades 3 and 6. CBM-W has only been investigated in two studies with divergent results using G-theory. Language structural differences also prevent the unreflected transfer of evaluation measures from one language to another: While the English orthography has a deep phoneme-grapheme correspondence, German has a more complex morphemic structure than English, which affects word length. German also has more complex rules for capitalization and punctuation (commas). Due to these linguistic differences, it is important to go beyond existing English-language studies to determine the optimal conditions for CBM-W in German.

The two central questions for the practical application of CBM-W, which we will address in the planned paper, are:

- (1) Which factors contribute to intraindividual variability in CBM-W, and to what extent?
- (2) Under which minimum measurement conditions does CBM-W achieve sufficient reliability for relative and absolute decisions?

The following hypotheses precede the data analyses:

- (1) In grade 3, prompts play a larger role, meaning that the facet story starter explains more variance than in grade 6. These differences are likely related to the fact that grade 3 children have less extensive vocabularies for certain topics and less world knowledge than grade 6 children. This, in turn, results in the younger children producing less text volume as they spend more time finding words and generating ideas. Thus, vocabulary size and vocabulary quality are likely to have less impact on the test score achieved as children get older.
- (2) Increasing writing time from 3 to 5 min positively affects measurement reliability at both grade levels, as reflected in higher G- and D-coefficients. In grade 6, this effect is even more positive since existing studies indicate that in lower grades, shorter writing samples are sufficient for reliable values, whereas, in higher grades, a longer writing time is appropriate to achieve adequate values.
- (3) Based on the observation that many children's achieved scores vary between measurement time points, it can be assumed that the D-coefficients differ significantly from the G-coefficients.

MATERIALS AND METHODS

Participants

Written expression curriculum-based measurement was conducted with a sample of third ($N = 128$) and sixth ($N = 118$) grade German-speaking students. Nine third grade classes and seven sixth grades classes from nine different schools participated in the study. The participating schools were spread over the German-speaking part of the canton of Fribourg (CH). Schools from both rural and urban areas participated in the study. The sample consisted of 71 girls (55.5%) and 57 boys (44.5%) in grade level 3 and 57 girls (48.3%) and 61 boys (51.7%) in grade level 6. A total of 119 students (48.4%) reported being multilingual, with 163 participants (66.2%) describing German as their first language. On average, the students were 8.8 years (SD 4.4) old in grade 3 and 11.8 years (SD 5.0) in grade 6. The active consent of the Education Directorate of the Canton of Fribourg, the school administrators, the class teacher, the parents and the child was a prerequisite for participation in the study.

Instrument

The instrument consists of five writing samples. The following story starters were used: "Last week I was allowed to take my pet to school when...", "I never believed in magic until Luke at school today...", "While walking on the beach, I discovered a stranded message in a bottle.", "Finally it worked, I invented the machine that...", "My feet are lifting off the ground. I'm flying!". These

five writing prompts were used in the same order for the third and sixth grades.

Procedure

The data collection was part of a larger study of writing fluency and its subcomponents.

Administration

The CBM-W samples were collected using a standardized implementation guide by teachers in the participating classes according to the usual standard for conducting CBM-W (Hosp et al., 2016). Students were given a sheet with the pre-printed story starter and lines to write on. They were told they had 1 min to think and then 5 min to write a story. After writing for 3 min, students were asked to mark with a cross the point to which they had written up to that point. The test administrator checked for accurate adherence to the time constraints. The students wrote the five writing samples within one school week.

The evaluations of the tests were done by trained students of special education. The training of the raters included an introduction to the scoring methods and the joint evaluation of several sample texts. There was the possibility to ask questions *via* an online forum during the data evaluation, which was actively used. No systematic checks were made to see if raters agreed with each other. In many other studies on writing assessment, training continues until high interrater reliability is ensured. Error variance attributable to the facet rater can thus be significantly reduced. The procedure chosen here realistically corresponds to the conditions under which CBM-W is implemented in school practice. The influence of the rater is presumably higher in school than in controlled studies, where many hours of rater training time are invested (Allen et al., 2019). Kim et al. (2017) also discuss that in a study examining factors influencing the reliability of a measurement method, it is preferable not to ensure a predetermined level of agreement between raters because the goal is to survey the influence of the facet rater under training conditions that are realistic in practice.

Scoring

This article focuses on four scoring methods: TWW, CWS, CIWS, and %CWS. These scoring methods include production-dependent measures (TWW, CWS), production-independent accuracy measures (%CWS) and accurate-production indices (CIWS) (Malecki and Jewell, 2003; Jewell and Malecki, 2005):

- Total Words Written (TWW): The number of written words separated by another by a space is counted. The words do not have to be spelled correctly (Espin et al., 2000).
- CWS: Fuchs and Fuchs (2007, 12) define CWS as follows: "A correct word sequence is one that contains any two adjacent, correctly spelled words that are acceptable within the context of the same to a native (English) speaker. The term 'acceptable' means that a native speaker would judge the word sequences as syntactically and semantically correct." Thus, the orthographic, semantic, and grammatical fit of what is written is assessed when

evaluating writing sequences. Correct writing sequences are marked with a carat between the two words. The evaluation of correct punctuation in English includes only the correct capital letter at the beginning of the sentence and the correct end mark at the end of the sentence. In German, we also evaluate the presence of necessary commas but not literal speech marks. In addition, it should be noted that in German, all nouns are capitalized, so capitalization is more complex than in English.

- **Correct Minus Incorrect Writing Sequences (CIWS):** Analogous to the correct writing sequences, incorrect writing sequences can also be evaluated. Between two words or a word and a punctuation mark, an incorrect sequence is then marked using an inverted carat if at least one of the two is incorrect in terms of orthography, semantics, or syntax. Missing elements (words or punctuation marks) in the present study were marked by two consecutive incorrect sequences. Subtracting the incorrect sequences from the correct ones yields an accurate production index, which incorporates writing fluency and accuracy (Jewell and Malecki, 2005).
- **Percentage of Correct Writing Sequences (%CWS):** This method – calculated as the percentage of correct sequences from the sum of correct and incorrect sequences – is independent of the amount of text written and is therefore considered a measure of accuracy (McMaster and Espin, 2007).

It should be noted that not every scoring method has proven to be equally reliable and valid at every grade level. While TWW is more suitable for younger students at the beginning of writing acquisition, CWS and CIWS are recommended for use around the third-grade level, but certainly for older students (McMaster and Espin, 2007; Saddler and Asaro-Saddler, 2013; McMaster et al., 2017; Romig et al., 2017; Payan et al., 2019).

Data Analysis

The statistical tests include analyses within the framework of G-theory (G-studies, D-studies). All calculations were performed separately for the third and the sixth grade for TWW, CWS, CIWS, and %CWS. The analyses were performed with the software G-String VI (Bloch and Norman, 2021)³, which is a graphical user interface for the operation of urGENOVA (Brennan, 2001). In the generalizability studies (G-studies), variance components were estimated for main and interaction effects of the facets student (facet of differentiation; between student differences), rater (differences across raters), and story starter (differences in performance across writing prompts). The resulting two-facet design is not fully crossed because the facet student is nested in raters.

Furthermore, it should be noted that the texts were assigned to the raters by class. This methodological aspect will be addressed in more detail in the discussion, but it is already mentioned here to better understand the data. The facet rater thus also includes the differences between different classes, which is why this facet is labeled rater/class in the results tables.

³https://github.com/G-String-Legacy/G_String/releases/tag/1.0.1/gstring_25.jar

The G-studies are calculated separately for the scoring methods TWW, CWS, and CIWS for 3 and 5 min of writing. Studies that also integrate duration of assessment as a facet must always collect student performance per minute (e.g., words read correctly per minute, math problems solved correctly per minute), since otherwise, the variance explained is simply a sign of more items solved in more time [see, e.g., Christ et al. (2005) and Keller-Margulis et al. (2016a)]. However, in the current study, student performance was not recorded after every minute but only after 3 vs. 5 min. The only scoring method for which time (differences in writing performance due to writing time) can be integrated as a facet in the G-study is the production-independent scoring procedure %CWS. This results in a three-facet design with the corresponding interactions.

In G-theory, negative variance components may occur. If these are small, they are usually set to zero (Stumpp and Großmann, 2009; Bloch and Norman, 2012; Briesch et al., 2014). In the present study, negative variance components are replaced by zero following this suggestion but are marked in the tables (*). To address research question 2, decision studies (D-studies) were subsequently conducted in G-string. These indicate how generalizability and dependability coefficients change when the measurement conditions vary (Briesch et al., 2014). Reported are both types of coefficients for one to five writing samples with 3- or 5-min writing time.

RESULTS

The descriptive results for all scoring methods and both grade levels are shown in **Table 1**. There is an increase in mean performance between the 3rd and 6th-grade levels for all scoring methods and through the increase in writing time.

Results of the G-Studies

The G-studies addressed the question of which factors contribute to the variability of the evaluated scoring methods and to what extent. **Table 2** documents the variance components for TWW, CWS, and CIWS in grades 3 and 6 for 5-min writing samples. The corresponding results for 3-min writing samples are similar to those presented here. They can be found in **Supplementary Table 1**. Obviously, the facet student explains the most variance for all scoring methods in the third and sixth grades. For 5-min writing samples, between 45% (CWS grade 3) and 64% (CIWS grade 6) turn out to be between-student differences. The rater/class facet also explains a significant portion of the variance, between 7 and 24%. The influence of story starter and the interaction story starter \times rater is extremely small in both grade levels and across all scoring methods, with a maximum of 3% variance explanation. Residual variance (i.e., non-systematic error) amounts to between 20 and 43%, whereby CIWS in grade 3 stands out due to a high proportion of error variance.

In the G-study for %CWS, time was included as a facet. **Table 3** shows that also, in this case, the facet of differentiation (student) explains a considerable proportion of the variance: about 35% for 3rd grade and 60% for 6th grade. Duration of Assessment (time) does not explain any variance at either grade level (0.02% each),

TABLE 1 | Descriptive statistics (M and SD).

		Grade 3				Grade 6			
		3 min		5 min		3 min		5 min	
		M	SD	M	SD	M	SD	M	SD
Probe 1	TWW	15.79	7.59	27.87	11.90	31.17	10.49	52.71	16.60
	CWS	7.02	4.81	12.60	7.61	24.29	10.68	40.44	17.61
	CIWS	−4.32	7.29	−7.15	11.15	12.09	14.66	19.34	23.84
	% CWS	39.29	18.53	39.00	15.86	65.70	17.91	64.82	16.99
Probe 2	TWW	18.96	8.02	31.52	13.22	32.29	11.74	53.15	18.86
	CWS	8.98	5.85	14.35	8.40	25.48	12.80	41.97	20.87
	CIWS	−4.31	8.98	−8.31	13.68	12.83	16.27	21.39	25.69
	% CWS	40.65	19.73	40.10	18.02	65.35	18.26	64.98	16.90
Probe 3	TWW	19.77	8.80	33.73	13.41	34.43	11.50	56.97	18.33
	CWS	9.56	5.99	15.54	9.12	27.30	12.47	44.85	20.67
	CIWS	−3.91	9.33	−8.35	14.95	14.37	17.11	23.00	28.00
	% CWS	42.61	19.98	40.34	17.26	67.14	18.36	65.96	17.65
Probe 4	TWW	19.30	9.02	33.60	14.16	34.28	13.90	57.96	20.78
	CWS	8.54	6.06	15.02	9.75	26.42	13.59	44.64	21.03
	CIWS	−5.70	9.30	−9.45	14.98	12.95	17.78	21.63	27.25
	% CWS	36.71	19.63	37.88	18.31	65.45	20.69	64.80	18.80
Probe 5	TWW	20.22	9.51	35.05	15.12	35.26	13.27	59.80	20.34
	CWS	10.10	6.99	16.59	10.85	27.48	13.68	46.89	22.19
	CIWS	−3.79	9.05	−8.54	14.60	13.46	17.72	22.86	29.60
	% CWS	41.50	19.57	39.60	17.17	64.80	19.24	64.51	18.93

TABLE 2 | Results of the G-studies for 5-min writing samples for TWW, CWS, and CIWS.

Facet	Grade 3		Grade 6	
	s ²	% s ²	s ²	% s ²
Results for TWW				
Rater/Class	48.33	24.78	46.36	12.46
Student (nested in Rater/Class)	96.35	49.41	213.38	57.36
Story starter	6.72	3.44	9.02	2.42
Rater/Class × Story starter	3.38	1.73	6.70	1.80
Residual	40.31	20.67	97.05	26.08
Total	195.09	100.03	372.51	100.12
Results for CWS				
Rater/Class	19.37	22.26	79.01	18.40
Student (nested in Rater/Class)	39.99	45.96	253.22	58.97
Story starter	2.08	2.39	5.84	1.36
Rater/Class × Story starter	0.49	0.56	1.57	0.37
Residual	26.04	29.93	89.74	20.90
Total	87.97	101.10	429.38	100.00
Results for CIWS				
Rater/Class	13.61	7.08	93.97	12.87
Student (nested in Rater/Class)	92.94	48.32	472.02	64.63
Story starter	−0.39	0.00	1	0.14
Rater/Class × Story starter	2.16	1.12	−1.48	0*
Residual	83.64	43.48	163.36	22.37
Total	192.35	100.00	730.35	100.00

*Negative variance components were set to zero. The sum may differ from 100 due to rounding.

nor does Story Starter. Particularly informative for %CWS is the Student \times Story starter interaction, which contributes most to variance explanation for 3rd grade (44%) and still accounts for 25% for 6th grade.

Results of the D-Studies

When addressing the question “with how many writing samples of which duration and with which scoring procedures does CBM-W achieve sufficient reliability for relative and absolute decisions?” we arrived at different answers for the two investigated grade levels: The results indicate that for the 6th-grade level, more complex scoring measures are indicated for relative decisions, but for the 3rd-grade level already the production measure TWW, measured by two writing samples of 5 min, is sufficient to exceed the threshold for low-stakes decisions of 0.80 (Table 4). Also, for CWS, three 5-min writing samples for the 3rd grade reach the value of 0.82, while CIWS and %CWS turn out to be inappropriate for this grade level. The situation is different at the 6th-grade level: for %CWS two 5-min writing samples reach 0.81, for CIWS and CWS already, two 3-min writing samples also reach 0.81, and for TWW, two 5-min writing samples are indicated. If one sets a stricter threshold of 0.90 for high-stakes decisions, it can be reached for students in grade level 3 only from four 5-min writing samples for TWW. In grade 6, the most time-efficient approach for achieving a relative reliability coefficient >0.90 would be to collect four 3-min samples using CWS or CIWS. Thus, while for relative decisions, procedures can be identified that are sufficiently reliable for making pedagogical decisions, this is not true for absolute decisions: only in one case is a benchmark of 0.80 reached for low-stakes decisions, and that is at the 6th-grade level for four 3- or 5-min writing samples.

DISCUSSION

Procedures for universal screenings and progress monitoring pursue the goal of reliably and validly recording and documenting the individual learning developments of students over time economically. For this purpose, they require parallel tests that show high stability and consistent within-student performance over time. Observations from a previous study on CBM-W (Winkes and Schaller, 2022) revealed, in contrast to this requirement, significant intraindividual variability in the writing performance of German-speaking primary school children over a short-term data collection period. In the present study, we chose generalizability theory as the methodological framework both to address the question of the big sources of variability for CBM-W and to investigate the effects of this variability on the reliability of CBM-W in terms of relative (rank order) and absolute (criterion-referenced) decisions under different measurement conditions.

So, what are the major sources of variability in CBM-W? On the positive side, a substantial portion of variance can be attributed to students (between student differences), ranging from 36 to 65%, depending on grade level and scoring measures. In grade three, student variance explanation is lower than in

grade six, where children explain about 60% of the variance for all scoring methods. For the G-studies without the time facet, the second-largest source of variance is unsystematic error variance (20–43%), followed by rater/class (7–25%). For the production-independent scoring method %CWS, assessment duration could be integrated as an additional facet in the G-study. Here, student-story starter-interaction emerges as the main source of variability in grade 3 (44%), ahead of between-student differences (35%). For sixth-graders, the variance explained by student \times story starter was much lower, but still 25%. It is also revealing which factors do not turn out to be a big source of variability, which is the case for story starters, for example. Thus, the very small differences between grade 3 and grade 6 are not significant, and hypothesis 1 (story starter has a more important role in grade 3 than in grade 6) could not be confirmed.

Hypothesis 2 assumed that increasing the writing time would positively affect the G- and D- coefficients. This hypothesis is supported, but the differences in the reliability coefficients between 3 and 5 min writing times are small in many cases. As predicted in hypothesis 3, the D-coefficients, on the other hand, deviates significantly from the G-coefficients. While between two and four writing samples are sufficient for relative decisions to exceed the threshold of 0.80, it is not reached by the D-coefficients for absolute decisions with one single exception (%CWS in 6th grade with four texts).

Which Sources of Variability Can Be Optimized for Written Expression Curriculum-Based Measurement?

Compared to other performance domains, assessments in the area of writing generally suggest an increased intraindividual variability. This is probably due in part to the complex cognitive demands of the writing process and in part to the open-ended tasks used in writing assessments (Kent and Wanzek, 2016; Ritchey et al., 2016). In the present study, approximately 60% of the variance was explained by students for all scoring methods for sixth-grade children and somewhat less for third graders. Other studies that examined children's writing performance using generalizability theory, using conventional evaluation methods (e.g., holistic or analytic teacher ratings), consistently found lower variance explained by the facet “person” [e.g., 10% in the study of Bouwer et al. (2015); 38–46% in the study of Graham et al. (2016) and 23–48% in the study of Schoonen (2012)]. Thus, in this respect, CBM-W is not inferior to other forms of writing assessments, also indicated by Kim et al. (2017).

The role of the writing prompt has been investigated for CBM-W primarily in the context of studies of parallel test reliability. However, these studies are less informative when an idiographic frame of comparison is applied, as is the case for progress monitoring (Christ and Hintze, 2007; Christ et al., 2016). That is why G-theory can make a relevant contribution here as an alternative to CTT. The analyses of variance within the G-studies presented indicate that story starters as a facet hardly explain variance. This finding is congruent with existing studies of writing that used G-theory (Schoonen, 2012; Keller-Margulis et al., 2016a; Wilson et al., 2019). This result may be considered

TABLE 3 | Results of the G-study for %CWS with time as a facet.

Facet	Grade 3		Grade 6	
	s^2	% s^2	s^2	% s^2
Rater/Class	26.46	7.62	23.17	6.77
Student (nested in Rater/Class)	124.34	35.83	207.90	60.78
Story starter	0.81	0.23	−0.57	0*
Time	0.08	0.02	0.22	0.02
Rater/Class × Story starter	0.90	0.25	2.10	0.61
Rater/Class × Time	0.21	0.06	−0.08	0*
Student × Story starter	153.02	44.09	86.36	25.25
Student × Time	−1.36	0*	0.49	0.14
Story starter × Time	0.61	0.17	−0.16	0*
Rater/Class × Story starter × Time	0.08	0.02	0.032	0.09
Residual	40.50	11.67	22.14	6.47
Total	347.01	99.96	342.70	100.13

*Negative variance components were set to zero. The sum may differ from 100 due to rounding.

positive in terms of the practical utility of CBM-W in that as many different story starters as desired can be used by teachers. The story starters do not differ systematically in terms of difficulty.

However, of great practical importance for using CBM-W is the interaction between student and story starter which proved to be a large source of variability when estimating the variance components for the scoring method %CWS. It explained 44% of the variance for the younger children (grade 3) and still 25% for the older children (grade 6). This result is in line with other studies on writing assessment, in which this effect also explained a very significant part of the variance (Schoonen, 2012; Bouwer et al., 2015; Graham et al., 2016). The question arises whether this effect in the mentioned studies is due to the combination of tasks of different genres or whether it also exists within one genre. Bouwer et al. (2015) used 12 texts (3 texts in each of four different genres), which were written at three different data collection points. They were able to show that generalizability of children's writing performance between different genres is not warranted (see also Graham et al., 2016). Writing assessments must therefore either include multiple texts of different genres or the interpretation of their results must be narrowed specifically to the genre used. However, the person × task interaction effect persists even within the same genre, as demonstrated by both Bouwer et al. (2015) and Kim et al. (2017). Specific to CBM-W, results to date have been inconsistent. While Kim et al. (2017) documented a large student × task interaction (both within the narrative genre and within expository genre), one did not occur in Keller-Margulis et al. (2016a). Our results support the assumption that it is not the individual story starters *per se* that contribute to variability but rather that children respond differently to tasks. As a possible explanation, it has been suggested that children's background knowledge and experiences differ concerning different writing tasks (Schoonen, 2005; Kim et al., 2017). Since CBM story starters are usually designed to accommodate the child's background experience (Hosp et al., 2016), this reasoning is not completely convincing. The story starters are very open in their formulations and allow the students to make associations in different directions, which is

why the world and background knowledge in a specific area should hardly carry any weight, especially since the content of the story is not the subject of the evaluation, but purely formal linguistic aspects are assessed. Therefore, supplementary explanations for the marked interaction effect between a person and a story starter should be considered. We suspect that, especially for younger children, the specific conditions of the writing assessments might have a significant influence, such as the time of day (morning, afternoon), whether the texts were written before or after recess, and which subjects were taught before, and so on. Furthermore, in the writing domain, motivational processes are considered to be of great importance. It is expected that children's personal and situational interests may vary with different writing stimuli and on different occasions (Troia et al., 2012). The influence of external conditions (e.g., time of day) could be included as an additional facet in future studies to verify this hypothesis.

For methodological reasons, the duration of the writing sample could only be integrated into the G-studies for %CWS. It did not explain any variance here, which is also consistent with the results of Keller-Margulis et al. (2016a), who investigated the influence of this facet on the generalizability of CBM-W more systematically. Thus, in the grade levels studied here, there is no evidence that intraindividual variability in the context of CBM-W is caused by the shortness of the writing sample and could be substantially reduced by longer writing samples. As described above, however, the duration of assessment could play a role in older students' writing (Weissenburger and Espin, 2005; Espin et al., 2008; Campbell et al., 2013).

Discussing the role of the rater is difficult for the current study because the raters were assigned by class and thus confounded with class (see below). Both together turn out to be variance components with a significant influence, explaining up to 25% of the variance. Whether differences between raters or between the performance of different classes in different schools manifest themselves here cannot be decided based on the present results and should thus be addressed in further research. However, we cannot exclude – also due to the somewhat more complex scoring

TABLE 4 | Results of the D-studies for TWW, CWS, CIWS, and %CWS.

Reliability coefficients for TWW					
Grade	<i>n</i> probes	Relative decisions (G-coefficient)		Absolute decisions (D-coefficient)	
		3 min	5 min	3 min	5 min
3	1	0.62	0.71	0.46	0.49
	2	0.76	0.82	0.55	0.57
	3	0.83	0.88	0.59	0.60
	4	0.86	0.91	0.62	0.61
	5	0.89	0.92	0.63	0.62
6	1	0.63	0.69	0.50	0.57
	2	0.77	0.81	0.61	0.67
	3	0.84	0.87	0.66	0.72
	4	0.87	0.90	0.68	0.74
	5	0.89	0.92	0.70	0.76
Reliability coefficients for CWS					
Grade	<i>n</i> probes	Relative decisions (G-coefficient)		Absolute decisions (D-coefficient)	
		3 min	5 min	3 min	5 min
3	1	0.52	0.61	0.40	0.45
	2	0.68	0.75	0.51	0.54
	3	0.77	0.82	0.56	0.58
	4	0.81	0.86	0.58	0.60
	5	0.84	0.88	0.60	0.61
6	1	0.68	0.74	0.53	0.59
	2	0.81	0.85	0.62	0.66
	3	0.87	0.89	0.65	0.69
	4	0.90	0.92	0.67	0.71
	5	0.92	0.93	0.68	0.72
Reliability coefficients for CIWS					
Grade	<i>n</i> probes	Relative decisions (G-coefficient)		Absolute decisions (D-coefficient)	
		3 min	5 min	3 min	5 min
3	1	0.42	0.53	0.40	0.48
	2	0.59	0.69	0.55	0.62
	3	0.68	0.77	0.64	0.69
	4	0.74	0.82	0.69	0.73
	5	0.78	0.85	0.73	0.75
6	1	0.68	0.74	0.60	0.65
	2	0.81	0.85	0.70	0.73
	3	0.87	0.90	0.74	0.76
	4	0.90	0.92	0.76	0.78
	5	0.92	0.94	0.77	0.79
Reliability coefficients for %CWS					
Grade	<i>n</i> probes	Relative decisions (G-coefficient)		Absolute decisions (D-coefficient)	
		3 min	5 min	3 min	5 min
3	1	0.39	0.42	0.36	0.38
	2	0.56	0.59	0.50	0.52
	3	0.66	0.68	0.57	0.59
	4	0.72	0.74	0.62	0.64
	5	0.76	0.78	0.65	0.67
6	1	0.66	0.68	0.61	0.63
	2	0.79	0.81	0.72	0.74
	3	0.85	0.86	0.77	0.79
	4	0.88	0.89	0.80	0.81
	5	0.90	0.91	0.82	0.83

rules for CBM-W in German – that the person evaluating has a relevant impact on the accuracy of the measurements.

Implications for the Use of Written Expression Curriculum-Based Measurement as a Screening and Progress Monitoring Tool

Conclusions for the use of CBM-W in practice can be drawn primarily from the D-studies. It should be noted that these only shed light on the aspect of reliability and must be supplemented for an overall conclusion by findings on the validity of the various scoring methods in different grades (McMaster and Espin, 2007; Romig et al., 2017). If we look only at the reliability results, we should distinguish between the use of CBM-W in the context of universal screenings and progress monitoring. These are two quite different tasks, but ideally, CBM-W should be suitable for both purposes (Payan et al., 2019).

Screenings whose goal is to identify the weakest writers in a group (Dunn, 2020) are typical contexts for relative decisions based on subjects' rankings. For this reason, G-coefficients are informative here if the group (e.g., class or students at the same level) rather than an external benchmark is used as a reference. It has already been shown in previous studies that the standard procedure, namely the collection of a single writing sample of 3 min, is not suitable to achieve sufficient reliability >0.80 (or even >0.90) (Keller-Margulis et al., 2021). Rather, depending on the grade level and scoring method, the evaluation of two to four 3-min writing samples is necessary for this purpose. Increasing the writing time leads in some constellations to the fact that fewer writing samples must be collected, but the total effort does not necessarily decrease. For example, in grade 3, relative reliability >0.80 is achieved with CWS by four 3-min samples (=12 min of writing time) or by three 5-min samples (=15 min of writing time). Accordingly, the feasibility and time-consuming nature of CBM-W as a universal screening tool is the main reason CBM-W is rarely implemented in practice (Payan et al., 2019). On the other hand, it must be stated that there are currently no alternatives for economical, reliable, and valid procedures to detect at-risk children in the area of writing in the context of universal screenings (Saddler and Asaro-Saddler, 2013). This underlines the need to understand more precisely the factors influencing the measurement accuracy of CBM-W and thus be able to optimize the procedure. Also, it should be reconsidered whether feasibility could be improved by reducing the frequency of screenings. It is recommended to conduct a writing screening three times a year with all students (Hosp et al., 2016; Traga Philippakos and FitzPatrick, 2018). However, Keller-Margulis et al. (2016a) found little within-year variance in student growth across different measurement points in the year in their study and therefore suggest limiting oneself to a single screening per year in the fall.

G-Theory provides an additional reliability coefficient in the form of the dependability coefficient. The D-coefficient focuses on the level of performance, regardless of rank. It is thus preferable for progress monitoring, in which students are compared with their performance over time (Fan and Hansmann,

2015). Concerning this intended use of CBM-W, we can conclude that the present analyses indicate that CBM-W is not sufficiently reliable – at least in German and in the grade levels studied – to be recommended for progress monitoring. For a single writing sample of 5 min duration, the highest D-coefficient in level 3 is 0.49 (TWW), and in grade 6 is 0.65 (CIWS) and fails to achieve a reliability >0.80 . Even by using multiple writing samples – which would be impractical for weekly assessments anyway – only one case (%CWS in grade 6 with four measurements of 3 or 5 min each) succeeds in achieving sufficient reliability. This result is, in fact, disappointing, but it reflects well our initial observation.

Limitations and Future Research

Finally, some methodological aspects should be discussed, which can be optimized in future studies by simple modifications. Reference has already been made to assigning children's texts to the raters, which leads to difficulties in interpreting the results. Texts were distributed to raters class by class. As a result, the facet "rater" is mixed with the factor class, and it is impossible to separate both factors' influence. Puranik et al. (2014) found significant differences between classes in writing instruction and the amount of time students spent on school writing activities in a study of kindergarten classes. This was reflected in a high variation in spelling and writing skills at the class level. This study also raises the possibility of a substantial influence of the "class" level on student performance. In future studies, children's texts should not be presented to raters on a class-by-class basis but should be randomized. Moreover, Bloch and Norman (2012) point out that it is also problematic when the same rater is involved in multiple subject ratings because rater variance is confounded with subject variance. Thus, if G-theory is used in the context of CBM, where there are usually always multiple samples of student performance, then randomization between tests and raters should continue consistently so that different raters evaluate different samples of a child.

A second possibility for optimization concerns the facet time. Only by including this facet in the %CWS method could the interesting interaction between student and story starter be uncovered. To consider the writing time as a facet for the other scoring methods, a marking in the text (or the change of pens) would be necessary (Christ et al., 2005; Keller-Margulis et al., 2016a) after every minute of writing time. Consideration of time is also reasonable in future G-studies of CBM-W because we still know too little about at what grade level and how great an increase in writing time is beneficial and therefore indicated.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation

and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JW led the team in data collection and participated in analyzing the writing probes. Both authors contributed to the conception and design of the study, performed all the analyses, and wrote the manuscript.

REFERENCES

- Alamargot, D., and Chanquoy, L. (2001). *Through the Models of Writing*, 1st Edn. Dordrecht: Springer Netherlands.
- Allen, A. A., Jung, P. -G., Poch, A. L., Brandes, D., Shin, J., Lembke, E. S., et al. (2019). Technical adequacy of curriculum-based measures in writing in grades 1–3. *Read. Writ. Q.* 33, 1–25. doi: 10.1080/10573569.2019.1689211
- Bloch, R., and Norman, G. R. (2012). Generalizability theory for the perplexed: a practical introduction and guide: amee guide no. 68. *Med. Teach.* 34, 960–992. doi: 10.3109/0142159X.2012.703791
- Bloch, R., and Norman, G. R. (2021). *G_String_VI: User Manual*. Hamilton, ON: McMaster University. doi: 10.1007/SpringerReference_28001
- Bouwer, R., Béguin, A., Sanders, T., and van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Lang. Test.* 32, 83–100. doi: 10.1177/0265532214542994
- Brennan, R. L. (2001). *Generalizability Theory. Statistics for Social and Behavioral Sciences Ser.* New York, NY: Springer, doi: 10.1007/978-1-4757-3456-0
- Briesch, A. M., Swaminathan, H., Welsh, M., and Chafouleas, S. M. (2014). Generalizability theory: a practical guide to study design, implementation, and interpretation. *J. Sch. Psychol.* 52, 13–35. doi: 10.1016/j.jsp.2013.11.008
- Campbell, H. M., Espin, C. A., and McMaster, K. L. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Read. Writ.* 26, 431–452. doi: 10.1007/s11145-012-9375-6
- Cardinet, J. (1998). Von der klassischen testtheorie zur generalisierbarkeitstheorie : der beitrag der varianzanalyse. *Bildungsforschung Und Bildungspraxis: Schweiz. Z. Erziehungswiss.* 20, 271–288.
- Christ, T. J., and Hintze, J. M. (2007). “Psychometric considerations when evaluating response to intervention,” in *Handbook of Response to Intervention: The Science and Practice of Assessment and Intervention*, eds S. R. Jimerson, M. K. Burns, and A. M. VanDerHeyden (Heidelberg: Springer), 93–105.
- Christ, T. J., Johnson-Gros, K. N., and Hintze, J. M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: the generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychol. Sch.* 42, 615–622. doi: 10.1002/pits.20107
- Christ, T. J., Van Norman, E. R., and Nelson, P. M. (2016). “Foundations of fluency-based assessments in behavioral and psychometric paradigms,” in *The Fluency Construct: Curriculum-Based Measurement Concepts and Applications*, eds K. D. Cummings and Y. Petscher (New York, NY: Springer), 143–163.
- Deno, S. L. (1985). Curriculum-based measurement: the emerging alternative. *Except. Child.* 52, 219–232. doi: 10.1177/001440298505200303
- Deno, S. L. (2003). Developments in curriculum-based measurement. *J. Spec. Educ.* 37, 184–192. doi: 10.1177/00224669030370030801
- DESI-Konsortium (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch*. Frankfurt: Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI).
- Dockrell, J. E., Connelly, V., Walter, K., and Critten, S. (2015). Assessing children's writing products: the role of curriculum based measures. *Br. Educ. Res. J.* 41, 575–595. doi: 10.1002/berj.3162
- Dunn, M. (2020). “What are the Origins and Rationale for Tiered Intervention Programming?,” in *Writing Instruction and Intervention for Struggling Writers: Multi-Tiered Systems of Support*, ed. M. Dunn (Newcastle upon Tyne: Cambridge Scholars Publisher), 1–15.

ACKNOWLEDGMENTS

We would like to thank Ralph Bloch for his indispensable support in using G-String.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2022.919756/full#supplementary-material>

- Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., and Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *J. Spec. Educ.* 34, 140–153. doi: 10.1037/spq0000138
- Espin, C., Wallace, T., Campbell, H. M., Lembke, E. S., Long, J. D., and Ticha, R. (2008). Curriculum-based measurement in writing: predicting the success of high-school students on state standards tests. *Except. Child.* 74, 174–193. doi: 10.1177/001440290807400203
- Fan, C.-H., and Hansmann, P. R. (2015). Applying generalizability theory for making quantitative RTI progress-monitoring decisions. *Assess. Effect. Interv.* 40, 205–215. doi: 10.1177/1534508415573299
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *Sch. Psychol. Rev.* 33, 188–193.
- Fuchs, L. S. (2017). Curriculum-based measurement as the emerging alternative: three decades later. *Learn. Disabil. Res. Pract.* 32, 5–7. doi: 10.1111/ldrp.12127
- Fuchs, L. S., and Fuchs, D. (2007). *Using CBM for Progress Monitoring in Written Expression and Spelling*. Available online at: <https://files.eric.ed.gov/fulltext/ED519251.pdf> (accessed March 31, 2022).
- Fuchs, L. S., Deno, S. L., and Marston, D. (1983). Improving the reliability of curriculum-based measures of academic skills for psychoeducational decision making. *Diagnostic* 8, 135–149. doi: 10.1177/073724778300800301
- Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., and Slider, N. J. (2002). Moving beyond total words written: the reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *Sch. Psychol. Rev.* 31, 477–497.
- Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., and Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *Sch. Psychol. Rev.* 35, 435–450.
- Graham, S., Hebert, M., Paige Sandbank, M., and Harris, K. R. (2016). Assessing the writing achievement of young struggling writers. *Learn. Disabil. Q.* 39, 72–82. doi: 10.1177/0731948714555019
- Graham, S., and Perin, D. (2007). *Writing Next: Effective Strategies to Improve Writing of Adolescents in Middle and High Schools – A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.
- Hintze, J. M., Owen, S. V., Shapiro, E. S., and Daly, E. J. (2000). Generalizability of oral reading fluency measures: application of G theory to curriculum-based measurement. *Sch. Psychol. Q.* 15, 52–68. doi: 10.1037/h0088778
- Hooper, S. R., Swartz, C. W., Wakely, M. B., de Kruijff, R. E. L., and Montgomery, J. W. (2002). Executive functions in elementary school children with and without problems in written expression. *J. Learn. Disabil.* 35, 57–68. doi: 10.1177/002221940203500105
- Hosp, J. L., and Kaldenberg, E. (2020). “What is writing assessment for tiered decision making?,” in *Writing Instruction and Intervention for Struggling Writers: Multi-Tiered Systems of Support*, ed. M. Dunn (Newcastle-upon-Tyne: Cambridge Scholars Publisher), 70–85.
- Hosp, M. K., Hosp, J. L., and Howell, K. W. (2016). *The ABC's of CBM: A Practical Guide to Curriculum-Based Measurement. The Guilford Practical intervention in the Schools Series*, Second Edn. New York, NY: The Guilford Press.
- Jewell, J., and Malecki, C. K. (2005). The utility of CBM written language indices: an investigation of production-dependent, production-independent, and accurate-production scores. *Sch. Psychol. Rev.* 34, 27–44.
- Keller-Margulis, M. A., Mercer, S. H., and Matta, M. (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement:

- a comparison study. *Read. Writ.* 34, 2461–2480. doi: 10.1007/s11145-021-10153-6
- Keller-Margulis, M. A., Mercer, S. H., and Thomas, E. L. (2016a). Generalizability theory reliability of written expression curriculum-based measurement in universal screening. *Sch. Psychol. Q.* 31, 383–392. doi: 10.1037/spq0000126
- Keller-Margulis, M. A., Payan, A., Jaspers, K. E., and Brewton, C. (2016b). Validity and diagnostic accuracy of written expression curriculum-based measurement for students with diverse language backgrounds. *Read. Writ. Q.* 32, 174–198. doi: 10.1080/10573569.2014.964352
- Kent, S. C., and Wanzek, J. (2016). The relationship between component skills and writing quality and production across developmental levels. *Rev. Educ. Res.* 86, 570–601. doi: 10.3102/0034654315619491
- Kim, Y. -S. G., Schatschneider, C., Wanzek, J., Gatlin, B., and Al Otaiba, S. (2017). Writing evaluation: rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Read. Writ.* 30, 1287–1310. doi: 10.1007/s11145-017-9724-6
- Malecki, C. K., and Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychol. Sch.* 40, 379–390. doi: 10.1002/pits.10096
- McMaster, K. L., and Espin, C. (2007). Technical features of curriculum-based measurement in writing. *J. Spec. Educ.* 41, 68–84. doi: 10.1177/00224669070410020301
- McMaster, K. L., Shin, J., Espin, C. A., Jung, P. -G., Wayman, M. M., and Deno, S. L. (2017). Monitoring elementary students' writing progress using curriculum-based measures: grade and gender differences. *Read. Writ.* 30, 2069–2091. doi: 10.1007/s11145-017-9766-9
- National Center for Education Statistics (2011). *The Nation's Report Card: Writing 2011*. Available online at: <https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf> (accessed March 29, 2022).
- Nunnally, J. C. (1967). *Psychometric Theory*, 5th Edn. New York, NY: McGraw-Hill.
- Payan, A. M., Keller-Margulis, M. A., Burrige, A. B., McQuillin, S. D., and Hassett, K. S. (2019). Assessing teacher usability of written expression curriculum-based measurement. *Assess. Effect. Interv.* 45, 51–64. doi: 10.1177/1534508418781007
- Poch, A. L., Allen, A. A., Jung, P. -G., Lembke, E. S., and McMaster, K. L. (2021). Using data-based instruction to support struggling elementary writers. *Interv. Sch. Clin.* 57, 147–155. doi: 10.1177/10534512211014835
- Puranik, C. S., Al Otaiba, S., Sidler, J. F., and Greulich, L. (2014). Exploring the amount and type of writing instruction during language arts instruction in kindergarten classrooms. *Read. Writ.* 27, 213–236. doi: 10.1007/s11145-013-9441-8
- Ritchey, K. D., McMaster, K. L., Al Otaiba, S., Puranik, C. S., Kim, Y. -S. G., Parker, D. C., et al. (2016). "Indicators of fluent writing in beginning writers," in *The Fluency Construct: Curriculum-Based Measurement Concepts and Applications*, eds K. D. Cummings and Y. Petscher (New York, NY: Springer), 21–66.
- Romig, J. E., Miller, A. A., Therrien, W. J., and Lloyd, J. W. (2020). Meta-analysis of prompt and duration for curriculum-based measurement of written language. *Exceptionality* 29, 133–149. doi: 10.1080/09362835.2020.1743706
- Romig, J. E., Therrien, W. J., and Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *J. Spec. Educ.* 51, 72–82. doi: 10.1177/0022466916670637
- Saddler, B., and Asaro-Saddler, K. (2013). Response to intervention in writing: a suggested framework for screening. *Intervention, and Progress Monitoring. Read. Writ. Q.* 29, 20–43. doi: 10.1080/10573569.2013.741945
- Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Lang. Test.* 22, 1–30. doi: 10.1191/0265532205lt295oa
- Schoonen, R. (2012). "The validity and generalizability of writing scores: the effect of rater, task and language," in *Measuring Writing: Recent Insights into Theory, Methodology and Practices Studies in Writing*, Vol. 27, eds E. van Steendam, M. Tillema, G. Rijlaarsdam, and H. van den Bergh (Boston, MA: Brill), 1–22. doi: 10.5271/sjweh.3746
- Stumpp, T., and Großmann, H. (2009). "Generalisierbarkeitstheorie," in *Enzyklopädie der Psychologie Methodologie und Methoden Evaluation: Bd. 1. Grundlagen und statistische Methoden der Evaluationsforschung*, eds H. Holling and N.-P. Birbaumer (Göttingen: Hogrefe Verl. für Psychologie), 207–234.
- The National Commission on Writing in America's Schools and Colleges (2003). *The Neglected "R": The Need for a Writing Revolution*. Available online at: https://archive.nwp.org/cs/public/download/nwp_file/21478/the-neglected-r-college-board-nwp-report.pdf?x-r=pcfile_d (accessed March 29, 2022).
- Traga Philippakos, Z. A., and FitzPatrick, E. (2018). A proposed tiered model of assessment in writing instruction: supporting all student-writers. *Insights Learn. Disabili.* 15, 149–173.
- Troia, G. A., Shankland, R. K., and Wolbers, K. A. (2012). Motivation research in writing: theoretical and empirical considerations. *Read. Writ. Q.* 28, 5–28. doi: 10.1080/10573569.2012.632729
- Weissenburger, J. W., and Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *J. Sch. Psychol.* 43, 153–169. doi: 10.1016/j.jsp.2005.03.002
- Wilson, J., Chen, D., Sandbank, M. P., and Hebert, M. (2019). Generalizability of automated scores of writing quality in Grades 3–5. *J. Educ. Psychol.* 111, 619–640. doi: 10.1037/edu0000311
- Winkes, J., and Schaller, P. (2022). Lernverlaufsdiagnostik schreiben (LVD – Schreiben): reliabilität, validität und sensitivität für mittelfristige lernfortschritte im deutschsprachigen raum. *Vierteljahress. Heilpädagogik Ihre Nachbargebiete* 91, 1–26. doi: 10.2378/vhn2022.art22d
- Zheng, Y., and Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from assessing writing (2000–2018). *Assess. Writ.* 42:100421. doi: 10.1016/j.asw.2019.100421

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Winkes and Schaller. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Statistical Power of Piecewise Regression Analyses of Single-Case Experimental Studies Addressing Behavior Problems

Jürgen Wilbert^{1*†}, Moritz Börnert-Ringleb^{2†} and Timo Lüke^{3,4†}

¹ Research Methods and Diagnostics, Institute of Inclusive Education, University of Potsdam, Potsdam, Germany, ² Institute of Special Education, Leibniz University Hannover, Hanover, Germany, ³ Inclusive Education and Improvement of Instruction, University of Graz, Graz, Austria, ⁴ Research Center for Inclusive Education, Graz, Austria

OPEN ACCESS

Edited by:

Yvonne Blumenthal,
University of Rostock, Germany

Reviewed by:

Mack Burke,
Baylor University, United States
Kaiwen Man,
University of Alabama, United States

*Correspondence:

Jürgen Wilbert
juergen.wilbert@uni-potsdam.de

†ORCID:

Jürgen Wilbert
orcid.org/0000-0002-8392-2873
Moritz Börnert-Ringleb
orcid.org/0000-0003-3533-0993
Timo Lüke
orcid.org/0000-0002-2603-7341

Specialty section:

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

Received: 11 April 2022

Accepted: 20 June 2022

Published: 06 July 2022

Citation:

Wilbert J, Börnert-Ringleb M and
Lüke T (2022) Statistical Power
of Piecewise Regression Analyses
of Single-Case Experimental Studies
Addressing Behavior Problems.
Front. Educ. 7:917944.
doi: 10.3389/feduc.2022.917944

In intervention research, single-case experimental designs are an important way to gain insights into the causes of individual changes that yield high internal validity. They are commonly applied to examine the effectiveness of classroom-based interventions to reduce problem behavior in schools. At the same time, there is no consensus on good design characteristics of single-case experimental designs when dealing with behavioral problems in schools. Moreover, specific challenges arise concerning appropriate approaches to analyzing behavioral data. Our study addresses the interplay between the test power of piecewise regression analysis and important design specifications of single-case research designs. Here, we focus on the influence of the following specifications of single-case research designs: number of measurement times, the initial frequency of the behavior, intervention effect, and data trend. We conducted a Monte-Carlo study. First, simulated datasets were created with specific design conditions based on reviews of published single-case intervention studies. Following, data were analyzed using piecewise Poisson-regression models, and the influence of specific design specifications on the test power was investigated. Our results indicate that piecewise regressions have a high potential of adequately identifying the effects of interventions for single-case studies. At the same time, test power is strongly related to the specific design specifications of the single-case study: Few measurement times, especially in phase A, and low initial frequencies of the behavior make it impossible to detect even large intervention effects. Research designs with a high number of measurement times show robust power. The insights gained are highly relevant for researchers in the field, as decisions during the early stage of conceptualizing and planning single-case experimental design studies may impact the chance to identify an existing intervention effect during the research process correctly.

Keywords: single-case design, single case analysis, Monte-Carlo simulation, behavior problems, special education, research design, single-case experimental design

INTRODUCTION

While experimental group designs are the most common way of testing educational and psychological research hypotheses, single-case experimental designs (SCED) experienced a renaissance over the last decades (Smith, 2012). In intervention research, SCEDs are a vital way to gain insight into the causes of individual changes that yield high internal validity (Kratochwill et al., 2010; Shadish et al., 2015). Among others, SCEDs are commonly applied to examine the effectiveness of classroom-based interventions to reduce behavioral problems in schools. Several literature reviews of SCED behavioral intervention studies have been published in the past few years. For example, Briesch and Briesch (2016) summarize the findings of single-case research on 48 behavioral self-management intervention studies. Soares et al. (2016) synthesized results of 28 single-case studies focusing on the effect size of token economy use in classroom settings. More recently, Moeyaert et al. (2021) summed up the body of research on the effects of peer-tutoring on academic and social-emotional outcomes and included 46 single-case studies. Several additional examples of the application of SCED in similar fields can be identified (e.g., Busacca et al., 2015; Harrison et al., 2019). However, at the same time, there is no consensus on good design characteristics of SCED when dealing with count data. Moreover, specific challenges arise concerning appropriate approaches to analyzing behavioral SCED data.

This paper aims to clarify these questions by specifying which factors (hereafter design specifications) influence the chance (i.e., statistical test power) of detecting an intervention effect in a single-case behavioral intervention study. In addition, based on the results gained, we aim to provide recommendations for SCED or at least to identify criteria for a researcher to consider when planning a single-case study.

Design Recommendations for Single-Case Studies

The most basic structure of a SCED consists of time series measurements on one individual divided into two phases: Continuous measurements occur before the start of a specific event (phase A) and continuous measurements taken after the event, e.g., the manipulation of an independent variable (phase B). This design can be extended to numerous variations regarding the number and order of phases (e.g., ABAB or AB1B2B3) based on specific research questions and assumptions on the nature of the behavior and the resulting data (Nock et al., 2007). Following the experimental logic of counterfactual thinking, the data of phase A serve as a reference for what would have happened in phase B if no intervention had taken place. Therefore, the level and development in phase B are compared to the level and development in phase A.

Despite the usefulness and importance of such SCEDs in applied research, researchers have to find common ground on how many measurements and phases should be included in SCED. Kratochwill et al. (2013) provide an overview of single-case intervention research design standards developed by a panel of experts in SCED methodology. However, these important

design recommendations include only very general design specifications and do not consider the specific characteristics of the measured feature (scaling and distribution). In contrast, we hypothesize that recommendations should be different when the measurements are count data (e.g., problem or error frequencies, which are Poisson distributed) or standardized scales (e.g., T or Z test scores, which are Gaussian distributed). We also hypothesize that choosing a particular SCED design depends on several design specifications (see **Figure 1**): the initial problem intensity at the start of a study, the intervention effect's expected strength (a level or a slope effect), and an expected data trend.

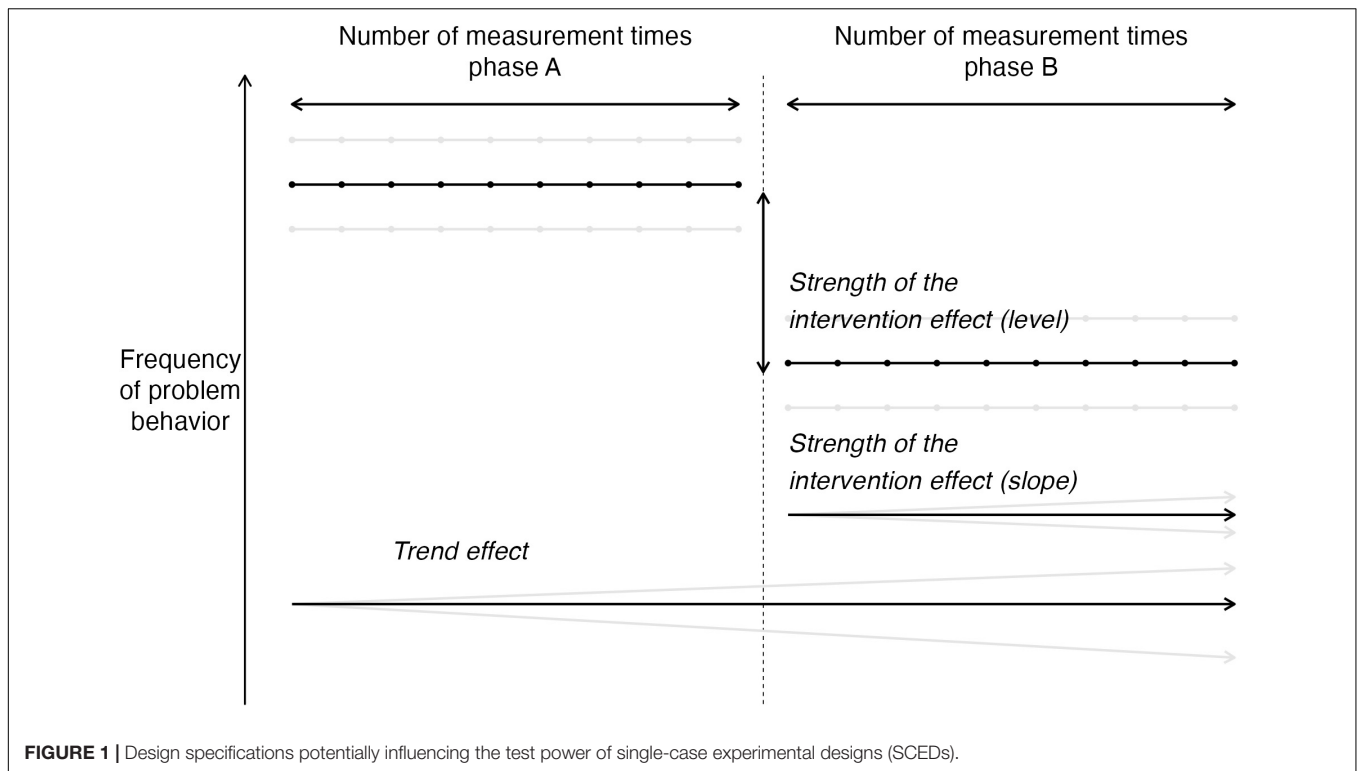
Single-Case Data Analyses

In addition to the design specifications, we also need to determine the method of data analysis since not all methods have the same sensitivity (or power). If someone decides to base the data analysis solely on visual inspection, one might recommend a different design than if the data analysis is based on a piecewise regression model.

Traditionally, single-case data have been analyzed through visual analysis (Parker and Brossart, 2003). Specifically, visual analysis is based on visual inspection of graphed time-series data where patterns related to level, trend, and overlapping/non-overlapping phases are evaluated to determine intervention effects (e.g., Parker and Vannest, 2012). Critics point out that visual analysis is overly subjective, vulnerable to misinterpretations due to data trends or outliers, and has less power (an increased type II error risk) compared to statistical analyses (Greenwald, 1976; Jones et al., 1978; Keppel, 1982; Matyas and Greenwood, 1990; Allison, 1992; Klapproth, 2018; Wilbert et al., 2021). There is evidence that agreement among multiple analysts and the consistency of their conclusions could be increased by using systematic protocols (Maggin et al., 2013; Wolfe et al., 2019).

Several statistical analysis techniques have been developed to overcome these critics throughout the last decades, either as a complement or a substitute for visual analysis. These procedures comprise overlapping indices (see Parker and Brossart, 2003) and "classical" statistical tests for comparing differences between groups like Student's *t*-tests and Mann-Whitney *U*-tests. These approaches have both benefits and significant limitations (e.g., not addressing autocorrelation and the existence of a trend throughout the data). Consequently, more complex statistical approaches have been applied to single-case data. These primarily include regression-based accounts (Huitema, 1986; Beretvas and Chung, 2008), randomization tests (Edgington and Onghena, 2007; Dugard et al., 2012; Heyvaert and Onghena, 2014), and mixed-effect models (Davis et al., 2013; Shadish et al., 2013; Moeyaert et al., 2014). These approaches address many of the former shortcomings like autocorrelation and the existence of a trend throughout the data (piecewise-regression models and randomization test), differentiate between immediate and continuous effects of an intervention (piecewise regression models), and allow the mutual analysis of several SCEDs (mixed models).

Many criteria considered in visual analysis are included and modeled in these more sophisticated statistical approaches (e.g.,



immediate and evolving intervention effects, data trends, data variability, complex phase contrasts). Other criteria specific to visual inspection may have to be investigated in more detail so they can be added to the statistical models explicitly (e.g., non-linearity of effects, outliers, lagged onset of intervention effects).

It is not easy to decide which approach is the “best” for analyzing single-case data. The underlying approaches to data analyses and statistics are fundamentally different: Piecewise regression analyses model data according to a complex theoretic model about the structure of single cases. Conversely, visual inspection relies on human expertise, pattern recognition, and intuition while overlap indices are targeted toward practitioners as an easy and accessible way to calculate effect sizes to validate their subjective judgment.

Based on the abovementioned arguments and studies, we consider piecewise regression models as one potentially appropriate and versatile approach among other alternatives. Notwithstanding, applying regression-based analyses (piecewise regression models and mixed models) comes with additional questions about the adequate distribution for modeling the dependent variable (more precisely, the error term) and the proper link function. Most implementations of regression analyses for SCED data are based on OLS estimators (e.g., Huitema and McKean, 2000) or generalized models with ML estimators based on Gaussian distributions (Ferron, 2002; Beretvas and Chung, 2008). While these estimators are adequate when the measured variable is continuous and normally distributed (e.g., a score in a standardized math test), they are less suitable for analyzing count data.

However, in single-case research, there are multiple types of dependent measures including count or frequency data. This is predominantly the case in SCEDs focusing on behavioral problems in schools: the dependent variable is often conceptualized as the frequency of a specific behavior within a certain period (e.g., disruptive or aggressive behavior). Frequencies are discrete numbers in nature; the Gaussian distribution models continuous values. Furthermore, frequencies can never be negative. Nevertheless, all negative numbers are modeled with a certain probability in a Gaussian probability density function. In line with this, Shadish and Sullivan (2011), in their overview of published SCED studies, argue:

Of particular interest is the fact that nearly all outcome variables were some forms of a count. Most parametric statistical procedures assume that the outcome variable is normally distributed. Counts are unlikely to meet that assumption and, instead, may require other distributional assumptions. In some cases, for example, the outcome is a simple count of the number of behaviors emitted in a session of a fixed length, which has a Poisson distribution (p. 979).

Binomial and Poisson distributions might be adequate alternatives. Binomial distributions display the probability of an outcome frequency given the number of events and the probability of an outcome for each event. Therefore, they are adequate for modeling count data and proportions (e.g., the frequency of behaviors). In cases where the occurring number of events is low, but the potential number of events is high, Poisson distributions are a viable alternative. These distributions depict a binomial distribution when the number of potential events approximates infinity, and the expected frequency of an

outcome (λ) is given. While a binomial distribution gives the probabilities of frequencies in the case of a finite exact number of possible occurrences, the Poisson distribution depicts the expected frequencies of an outcome when the number of possible occurrences approximates infinity. Such conditions are often met when behavioral data are measured. Consider, for example, a researcher investigating the occurrence of inappropriate behavior. At its extreme, a student might show inappropriate behavior at any second. At the same time, it is also realistically possible that no inappropriate behavior occurs at all.

Despite these arguments, piecewise Poisson-regression models are not widespread in SCED research. This depicts a potential limitation to existing studies as effects might not have been adequately identified as relying on flawed distributional assumptions impacts the power of the chosen analytical approach. In addition, the use of Poisson distributions in regression models as means of analyzing SCED has not been examined in detail. Insights into test power and alpha error rate are lacking. However, such insights might yield crucial additional information on the adequacy of the design specifications of SCED.

Study Aims

The present paper aims to investigate the test power of piecewise regression analyses for analyzing SCEDs with count data. Thereby, we aim to address the impact of essential design specifications of SCEDs on test power. More specifically, we examine the influence of the following aspects on the test power:

- (1) The initial frequency of the (problem) behavior,
- (2) The strength of the intervention effect,
- (3) The number of measurement times in phase A (baseline) and phase B (intervention),
- (4) The interaction between initial frequency, the strength of the intervention effect, and the number of measurement times,
- (5) The interaction of the number of measurement times in phase A and phase B, and the initial frequency of the behavior,
- (6) The presence of a trend in the data,
- (7) The interaction of a trend in the data, the strength of the intervention effect, and the number of measurement times.

Besides the test power, we will also report the alpha-error probabilities (type I errors) for all investigated conditions. Our regression approach will extend the piecewise regression model proposed by Huitema and McKean (2000) to include Poisson distributed dependent variables. These insights might depict an important orientation for deriving design principles of adequate SCED in the context of behavioral data.

MATERIALS AND METHODS

To answer the research questions mentioned above, we set up several Monte-Carlo simulation studies that focused on specific design specifications of SCEDs. The general idea behind such simulations is to generate a high number of random single-case

datasets with specified conditions (e.g., a specific intervention effect). Afterward, these datasets are analyzed (here, using a piecewise Poisson-regression model). Comparing the results of each analysis to the initial setup of the random case generates four results:

- (1) True-positive: The initial setup contained an intervention effect, and the analysis found a significant effect.
- (2) True-negative: The initial setup did not contain an intervention effect, and the analysis did not find a significant effect.
- (3) False-positive: The initial setup did not contain an intervention effect, and the analysis found a significant effect.
- (4) False-negative: The initial setup did contain an intervention effect, and the analysis did not find a significant effect.

The proportion of true positive results is the *power*, and the proportion of the false-positive results is the *alpha error probability* of a test for the given design specifications.

Data Simulation Rationale

The data simulation followed the rationale elaborated below. For any studies applying a Monte-Carlo approach, the validity of the findings and their relevance to practice depend on the characteristics of the data generated. Therefore, we paid particular attention to aligning the simulated data, if reasonable, with the reality of published SCED studies.

Phase Design

AB-Designs are the simplest form of a SCED comprised of a baseline (phase A) and an intervention phase (phase B). At the same time, AB depicts the building block for any multiple-phase design, and the multiple baseline design (MBD) – the most frequent SCED (Shadish et al., 2014). Therefore, we decided to choose an AB design as the underlying phase design of the simulated data.

Outcome Variable

We were particularly interested in analyzing intervention studies in which a teacher or researcher attempts to reduce a specific (problematic) behavior during classroom learning. Here, the target behavior is captured through systematic direct observations (e.g., Hintze et al., 2002; Lane and Ledford, 2014; Ledford et al., 2018), which are the “most widely used outcomes in single-case research” (Pustejovsky, 2018, p. 100). Thus, we used Poisson-regression models. The simulated data should represent count data (frequency of the observed behavior).

Initial Problem Behavior Frequency

Another potential factor influencing the test power and alpha-error probability of the analyses is the frequency of the dependent variable. The behavior of interest to the particular research question may be scarce (e.g., self-harming behavior during class) or widespread (e.g., disturbing behavior). Hence, the problem behavior frequency depends on the behavior of interest and the exact operationalization. Therefore, we decided to set up a

simulation where we vary the expected problem intensity starting with a low frequency of 5 to a high frequency of 30. These frequencies follow the mean baseline frequencies of adverse valence outcomes described in the overview of 303 published SCEDs provided by Pustejovsky et al. (2019, p. 24). In simulations where we did not focus on the relevance of behavior frequencies, we chose an expected behavior frequency of 15.

Number of Measurement Times in Phase A

A certain proportion of published SCED studies include fewer than three phase A measurement times (Pustejovsky et al., 2019). This contradicts both current recommendations (e.g., Kratochwill et al., 2013) and the basic requirements of regression methods. We simulated single-case data using a minimum of three measurements per phase following usual conventions (e.g., Hitchcock et al., 2014). Further, Pustejovsky et al. (2019) found that the number of phase A measurements was below 20 for the overwhelming majority of SCED studies. Most studies had between 2 and 15 phase A measurement times. Therefore, we set up a simulation varying the length of phase A between 3 and 19 measurements. In line with Smith (2012), who found an average of 10.2 phase A observations in their review of 400 published SCED studies, we used 10 phase A measurements for the other simulations.

Number of Measurement Times in Phase B

In addition to varying phase A (baseline) lengths, the number of measurement times in phase B (intervention) also varies, for example, due to the number of sessions of an implemented intervention. We, therefore, varied the number of measurement times (the length) of phase B in one simulation. Usually, the length of phase B exceeds the length of phase A. We took this into account by setting the minimum length of phase B to 10 measurements and the maximum to 50. We set 20 phase B measurements as a fixed value for the other simulations.

Intervention Effect

Another essential characteristic of SCED studies is the strength of the intervention effect (i.e., the reduction of the problem behavior). Most of the published research using SCEDs usually reports quite significant effects; however, it needs to be considered that this might also be due to a publication bias (Travers et al., 2016; Dowdy et al., 2022). In addition, the majority of the published SCED studies report different measures of effect sizes (such as overlap indices). Only a few studies report effect sizes associated with regression analysis. Therefore, it is difficult to derive an expected “mean” intervention effect from existing studies. We addressed this challenge by setting up a simulation with varying intervention effects employing the level effect between 20% and 80% problem reduction. We used a reduction of the dependent variable by 50% for the other simulations. In practice, behavior reductions of this magnitude are considered substantial (Vannest and Sallese, 2021, p. 17). We further assumed that, on average, no additional slope effect would be present in the data, but we included a slope effect for each case randomly drawn from a gaussian distribution with a mean of zero and a standard deviation of 10% of the initial problem behavior

frequency. We considered that an intervention does not exactly exert the same effects on every individual.

Trend Effect

Another common feature of single-case data is the presence of a trend effect in the data. This trend indicates an overall development in the problem behavior, which already appears in phase A (baseline) and is independent of the intervention. This trend might be positive (increasing the problematic behavior frequency across time) or negative (reducing the problematic behavior) and depends on many individual variables (e.g., additional support from home; negative peer influence; maturation). Therefore, we set up a simulation for positive and negative trend effects by varying the trend's strength between a decrease of 60% to an increase of 60% of the problem behavior frequency throughout all measurements. For all the other simulations, we included a random trend effect for each simulated single-case drawn from a gaussian distribution with a mean of zero and a standard deviation of 10% of the initial problem behavior.

Monte-Carlo Design

We conducted three simulations. Each simulation varied specific SCED specifications.

For simulation 1, we varied the intervention effect (4 iterations: -0.2 ; -0.4 ; -0.6 ; -0.8), the number of measurement times in phases A (9 iterations: 3, 5, 7, 9, 11, 13, 15, 17, 19), and the number of measurement times in phase B (7 iterations: 10, 15, 20, 25, 30, 40, 50), resulting in $4 \times 9 \times 7 = 252$ design conditions.

For simulation 2, we varied the initial frequency of the behavior (6 iterations: 5; 10; 15; 20; 25; 30), the intervention effect (4 iterations: -0.2 ; -0.4 ; -0.6 ; -0.8), and number of measurement times (6 iterations: 15, 21, 27, 33, 39, 45 where 1/3 of the measurements belong to phase A and 2/3 to phase B), resulting in $6 \times 4 \times 6 = 144$ design conditions.

For simulation 3, we varied the initial frequency of the behavior (6 iterations: 5; 10; 15; 20; 25; 30), the trend effect (5 iterations: -0.6 ; -0.4 ; 0; 0.4 ; 0.6), and number of measurement times (6 iterations: 15, 21, 27, 33, 39, 45 where 1/3 of the measurements belong to phase A and 2/3 to phase B), resulting in $6 \times 5 \times 6 = 180$ design conditions.

For each design condition within each simulation, 10,000 random single cases with the respective design specifications were generated (the generation algorithm below). Each case was analyzed with a piecewise Poisson-regression model (see below). The proportion of significant intervention effects in these analyses is the test power for the respective attributes for that design condition.

In a second step, another 10,000 random single-cases were created for each design condition. This time, the intervention effect was set to zero for all cases. Again, each case was analyzed with a piecewise Poisson-regression model. The proportion of significant intervention effects detected in these analyses is the design condition's alpha-error probability.

Preparatory tests have shown that we need a rather high number of 10 000 cases per variant to achieve a stable estimate. This is due to various random parameters and several interactions

TABLE 1 | Overview of the parameter settings and iterations (runs) for the three simulations.

Parameter	Simulation 1	Simulation 2	Simulation 3
Initial behavior frequency (<i>start</i>)	15	{5, 10, 15, 20, 25, 30}	{5, 10, 15, 20, 25, 30}
Phase A and B length (MT_A/MT_B)	$MT_A = \{3, 5, 7, 9, 11, 13, 15, 17, 19\}$ crossed ¹ with $MT_B = \{10, 15, 20, 25, 30, 40, 50\}$	$MT_{A+B} = \{15, 21, 27, 33, 39, 45\}$ with $MT_A = 1/3$ and $MT_B = 2/3$ of the length.	$MT_{A+B} = \{15, 21, 27, 33, 39, 45\}$ with $MT_A = 1/3$ and $MT_B = 2/3$ of the length.
Intervention effect (<i>level</i>)	{-0.2, -0.4, -0.6, -0.8}	{-0.2, -0.4, -0.6, -0.8}	-0.5
Trend effect ² (<i>trend</i>)	$\mathcal{N}(\mu = 0, \sigma^2 = 0.1 \times \frac{start}{MT_{A+B}})$	$\mathcal{N}(\mu = 0, \sigma^2 = 0.1 \times \frac{start}{MT_{A+B}})$	{-0.6, -0.4, 0, 0.4, 0.6} * $\frac{start}{MT_{A+B}}$
Slope effect ³ (<i>slope</i>)	$\mathcal{N}(\mu = 0, \sigma^2 = 0.1 \times \frac{start}{MT_B})$	$\mathcal{N}(\mu = 0, \sigma^2 = 0.1 \times \frac{start}{MT_B})$	$\mathcal{N}(\mu = 0, \sigma^2 = 0.1 \times \frac{start}{MT_B})$

Curly brackets depict iterations. ¹Crossed means that each iteration in MT_A is combined with each iteration in MT_B . ²A trend effect is a continuous change of the behavior frequency independent of the intervention effect and across all measurement times. ³A slope effect is a continuous change of the behavior frequency due to the intervention and across Phase B.

that go into the data generation algorithm (see section “Results” and Table 1).

The random cases were generated with the R package *scan* (Wilbert and Lüke, 2022). The same package was used for calculating the test power and alpha error probability. The source code for all analyzes is available as an online supplement to this paper¹.

Data Generation Algorithm

Firstly, a random single-case was created by calculating the expected behavior frequency (λ) for each measurement (i). The formula adapts a piecewise-regression model for single cases:

$$\lambda_i = start + level \times start \times phase_i + trend \times mt_i + slope \times phase_i \times (mt_i - MT_A) \quad (1)$$

where,

i = The index of a measurement.

$start$ = The initial problem frequency at the start of the study.

$phase$ = A variable with 0 for phase A and 1 for phase B measurements.

$level$ = The change of expected problem behavior frequency due to the intervention (e.g., -0.5 for a 50% reduction).

mt = The measurement time.

$trend$ = A trend effect leading to a change in problem behavior frequency for each measurement. Calculated by $\mathcal{N}(\mu = 0, \sigma^2 = 0.1 \times \frac{start}{MT_{A+B}})$.

$slope$ = A change of expected problem behavior frequency for each measurement that starts with the onset of phase B. For simulations 1 and 2 calculated by $\mathcal{N}(\mu = 0, \sigma^2 = 0.1 \times \frac{start}{MT_B})$.

MT_A = The number of measurement times of phase A.

Second, the observed values for each measurement y were drawn from a Poisson distribution with the expected probability:

$$P(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad (2)$$

Depending on the respective aim of the simulation, $start$, MT_A , MT_B (the total number of measurements – MT_A), $level$, and $trend$ were varied.

¹<https://osf.io/ys3a9/>

Figure 2 shows three corresponding examples of single cases.

Data Analyses Model

Each randomly generated case was re-analyzed with a piecewise regression model (Huitema and Mckean, 2000) adapted for Poisson distributed data:

$$\log(y_i) = \beta_0 + \beta_1 mt_i + \beta_2 phase_i + \beta_3 phase_i(mt_i - MT_A) + e_i \quad (3)$$

Table 2 shows an example of a piecewise Poisson-regression analysis for the first example case of Figure 2. Here, the level phase B effect is significant ($B = -1.18$, $p < 0.01$). As the original construction algorithm for that single case entailed an intervention effect, the result of this analysis is true-positive.

RESULTS

In the present study, we investigated how various specifications of SCEDs affect the statistical power of regression-based analyses assuming Poisson-distributed behavioral data. In addition, we focused on those design parameters that we believe are most frequently discussed and most likely to be influenced by researchers when planning a SCED. All figures in this paper are created with the software packages *ggplot* (Wickham, 2016) and *scplot* (Wilbert, 2022). All data and analyses are reproducible and made available on the project page (see text footnote 1).

Simulation 1: Intervention Effect and Number of Measurement Times in Phases A and B

First, we examined the statistical power as a function of the intervention effect and the number of measurement times in phases A and B. The initial frequency of the behavior is kept constant, and the trend- and slope effect sizes are randomly generated for each case with an expected value of zero (see Table 1).

Figures 3A–D depict the power (blue lines) and alpha-error probability (red lines) for all design conditions. The figures also include lines marking the usually recommended minimal power level of 80% and the maximum alpha-error probability of 5%.

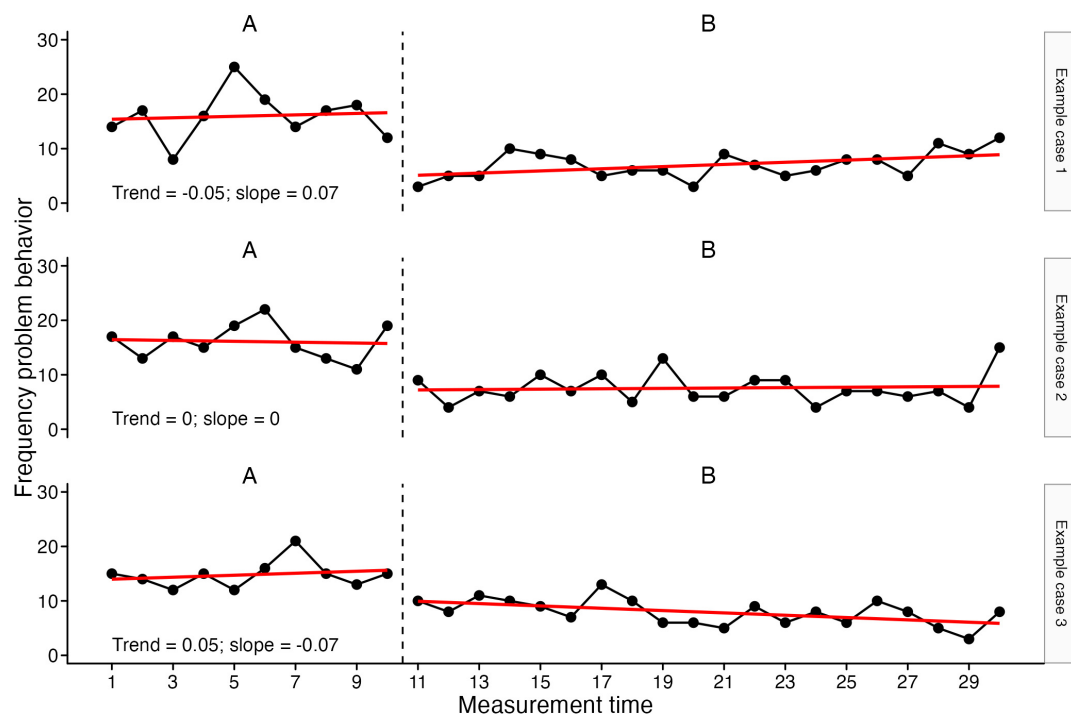


FIGURE 2 | Three random single cases based on exemplary design specifications.

TABLE 2 | Piecewise Poisson-regression table, for case 1 in **Figure 2**.

Parameter	B	2.5%	97.5%	SE	t	P
Intercept	2.73	2.38	3.05	0.17	15.81	< 0.01
Trend	0.01	-0.05	0.06	0.03	0.30	0.76
Level phase B	-1.18	-1.65	-0.71	0.24	-4.93	< 0.01
Slope phase B	0.02	-0.04	0.08	0.03	0.65	0.52

$\chi^2(3) = 54.38$; $p < 0.001$; AIC = 154.

The length of phase A is plotted on the x -axis (between 3 and 19). The shape of the dots describes the respective length of phase B (between 10 and 50). The facets (**Figures 3A–D**) refer to the strength of the intervention effect (between -0.2 and -0.8).

No relevant power is obtained for a small intervention effect (20% reduction of the problem behavior) regardless of the number of measurement times in phases A and B (**Figure 3D**). With a reduction of problem behavior by 40% at the beginning of phase B (**Figure 3C**), a significant power of more than 80% is only achieved with a large number of measurement times; more precisely, with 11 measurement times in phase A and ≥ 50 measurement times in phase B, as well as with 13 measurement times or more in phase A and ≥ 40 measurement times in phase B. If the intervention reduces the problem behavior by 60% (**Figure 3B**), sufficient power is achieved with designs of ≥ 5 measurement times in phase A and ≥ 15 measurement times in phase B, improving further with ≥ 11 measurement times in phase A. For designs with ≤ 10 measurement times in phase B, sufficient power is achieved only with ≥ 9 measurement times in phase A. With an 80% reduction of the problem behavior

with the intervention's start, statistical power is satisfactory across all design conditions (**Figure 3A**). In particular, with ≥ 15 measurement times in phase B, the probability of detecting an intervention effect is high regardless of the number of measurement times in phase A.

The alpha-error probability is stable at 5% across all design conditions.

Simulation 2: Initial Frequency of the Behavior, Intervention Effect, and Number of Measurement Times

Next, we consider the influence of the intervention effect size, the initial behavior frequency, and the total length of the design (see **Table 1** for a list of all parameters in this simulation). **Figure 4** shows the results and is analogous to **Figure 3**. The number of measurement time points ($1/3$ phase A and $2/3$ phase B) is plotted on the x -axis (between 15 and 45). The shape of the dots describes the strength of the intervention effect (between -0.2 and -0.8). The facets (**Figures 4A–F**) refer to the initial frequency of the behavior (between 5 and 30).

Both a very low initial behavior frequency and few measurement times lead to poor test power. Regardless of the other specifications of the design, small intervention effects (20% reduction of problem behavior at the beginning of phase B) cannot be detected reliably (**Figures 4A–D**, lines with crosses). When the intervention reduces the target behavior by 40% (lines with squares), sufficient power is achieved only when the initial behavior frequency and the number of measurement times are high (≥ 20 initial behavior frequency and ≥ 33 MT;

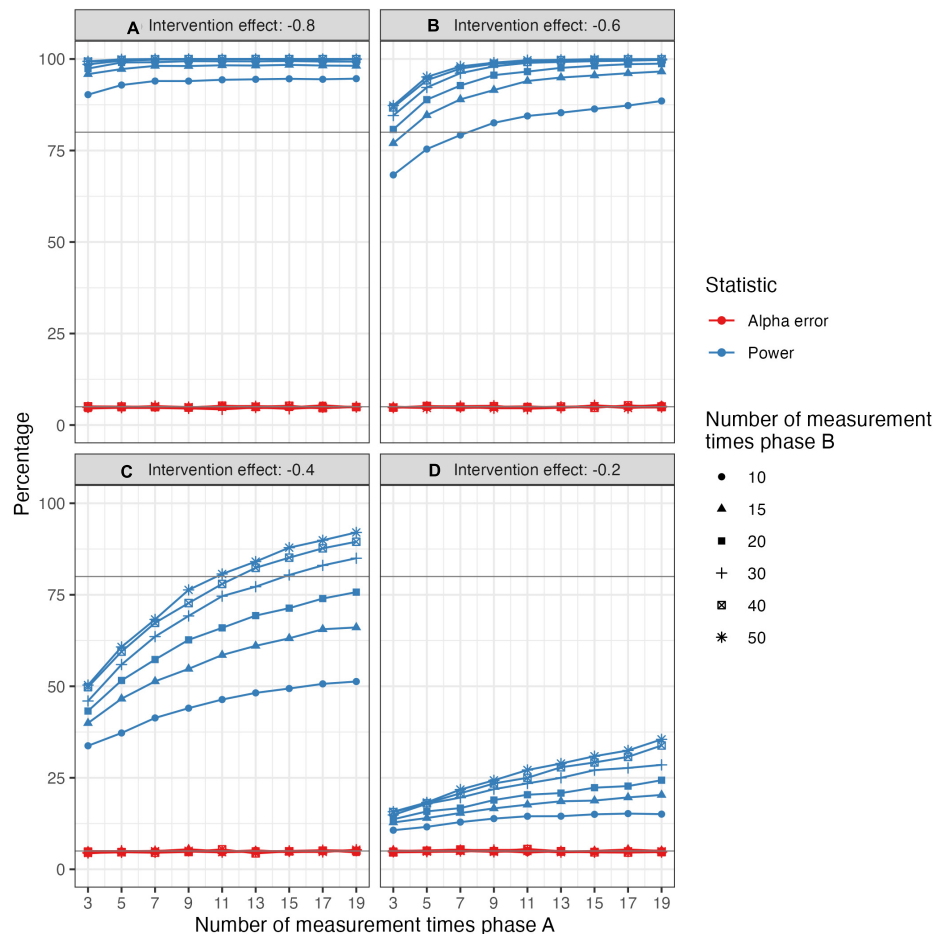


FIGURE 3 | Power and alpha error rates (line colour) for different intervention effect sizes (part) and measurement times per phase (dot shape and x-axis) (simulation 1).

≥ 25 initial behavior frequency and ≥ 27 MT). Large intervention effects such as an 80% reduction in problem behavior can be reliably detected at initial behavior frequencies of ≥ 10 . For medium-level effects of the intervention (60% reduction; **Figure 4**, lines with triangles), a sufficient power depends on the combination of the other conditions: If the initial behavior frequency is ≥ 20 , sufficient power is reliably achieved. With ≥ 30 measurement time points, sufficient power is achieved even with an initial frequency of 10 or 15. With an initial behavior frequency of 5, on the other hand, even a large number of measurement times no longer helps to achieve sufficient power.

The alpha-error probability is stable at 5% for all design conditions.

Simulation 3: Initial Frequency of the Behavior, Data Trend, and Number of Measurement Times

Finally, we would like to consider in more detail the interplay of the initial behavior frequency, the number of measurement times, and the data trend (see **Table 1** for a list of all parameters in this

simulation). **Figure 5** depicts the results and is built analogous to the previous figures. The number of measurement time points (1/3 phase A and 2/3 phase B) is plotted on the x-axis (between 15 and 45). The shape of the dots describes the strength of the trend effect (between -0.6 and 0.6). The facets (**Figures 5A-F**) refer to the initial frequency of the behavior (between 5 and 30).

A data trend of 60% reduction in the problem behavior frequency throughout the study (**Figure 5**, lines with circles) strongly reduces the test power for all design conditions. Only exceptionally high initial levels of the problem behavior (≥ 25) and large numbers of measurement times (≥ 33 ; **Figure 5E**) show a power level $\geq 80\%$. In cases with a weaker, negative data trend ($\geq -40\%$), this problem is no longer observed (lines with triangles), and the power is comparable to designs without a trend.

The effect of the initial behavior frequency on the test power described in section “Simulation 3: Initial Frequency of the Behavior, Data Trend, and Number of Measurement Times” can be similarly identified here: Only for designs with an initial behavior frequency ≥ 15 sufficient power is achieved for most cases. Especially in **Figures 5C-E**, the interaction between all

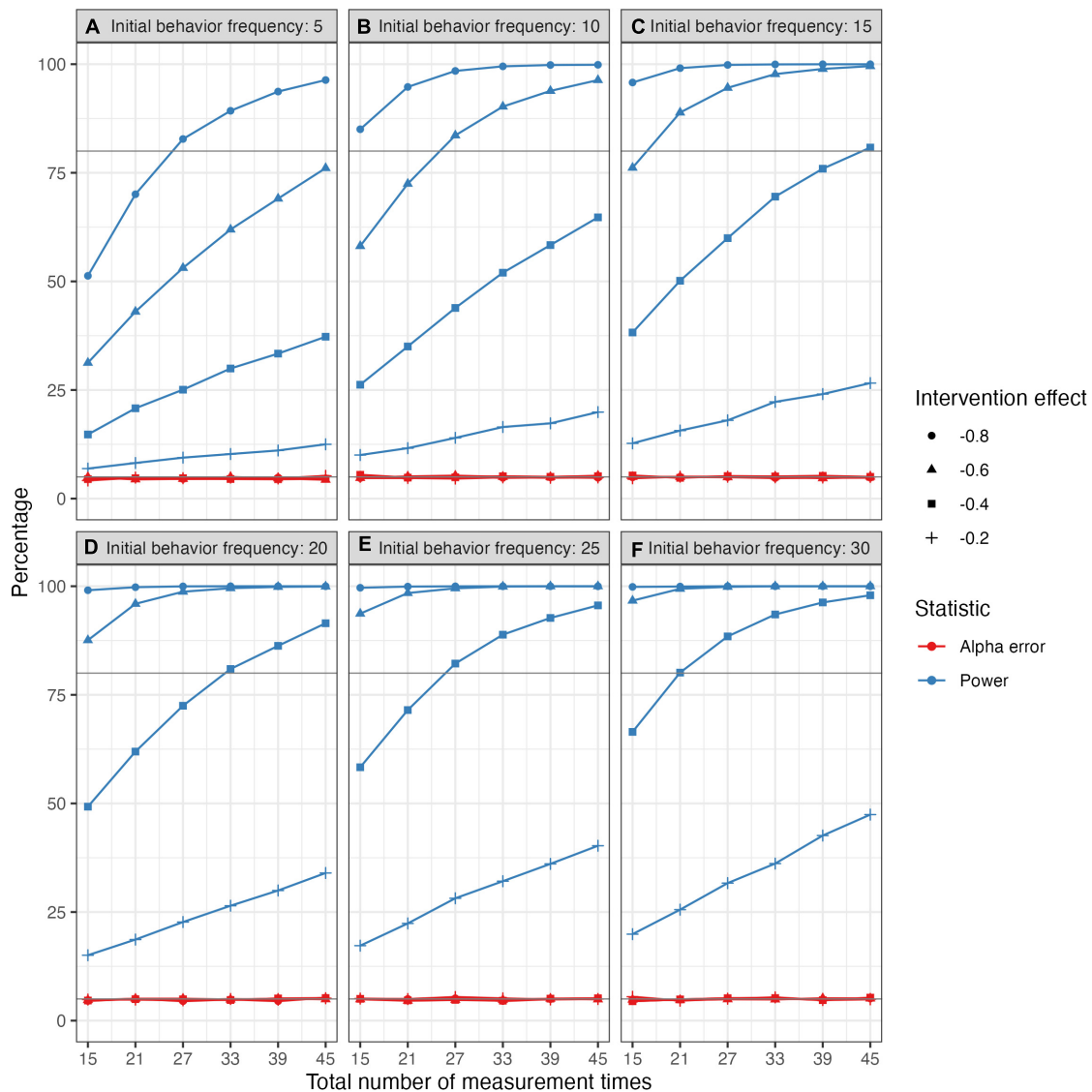


FIGURE 4 | Power and alpha error probability (line colour) for different intervention effect sizes (dot shape), initial behavior frequency (part), and number of measurement times (x-axis) (simulation 2).

three parameters becomes apparent: While sufficient power is not achieved for designs with a substantial negative data trend, the detection rate is acceptable for increasing (or stable) problem behavior (≥ 0) and designs with ≥ 27 measurement times. In cases with a high initial behavior frequency (≥ 25 ; **Figure 5E**), the power approaches 100% quite rapidly.

Again, the alpha-error probability is stable at 5% for all design conditions.

DISCUSSION

The goal of the paper at hand was to shed light on the usefulness of applying piecewise Poisson-regression models (in terms of statistical power) to analyze single-case data under varying design

specifications. Specifically, we investigated the influence of phase length, intervention effect size, initial frequency of the dependent variable, and the size of a trend effect on test power.

Overall, the results of the conducted simulations indicate that Poisson-regressions have a high potential of identifying (i.e., a test power of 80% or higher) intervention effects. However, at the same time, the test power was low under specific conditions. Hence, following our theoretical assumptions, test power seems to be related to the specific design specifications of the SCED study. The alpha-error probability was 5% for all conditions, even with very strong trend effects. The insights gained are highly relevant for researchers in the field, as design decisions during the early stage of conceptualizing and planning SCED studies might impact the overall potential of correctly identifying an existing intervention effect. Our results might

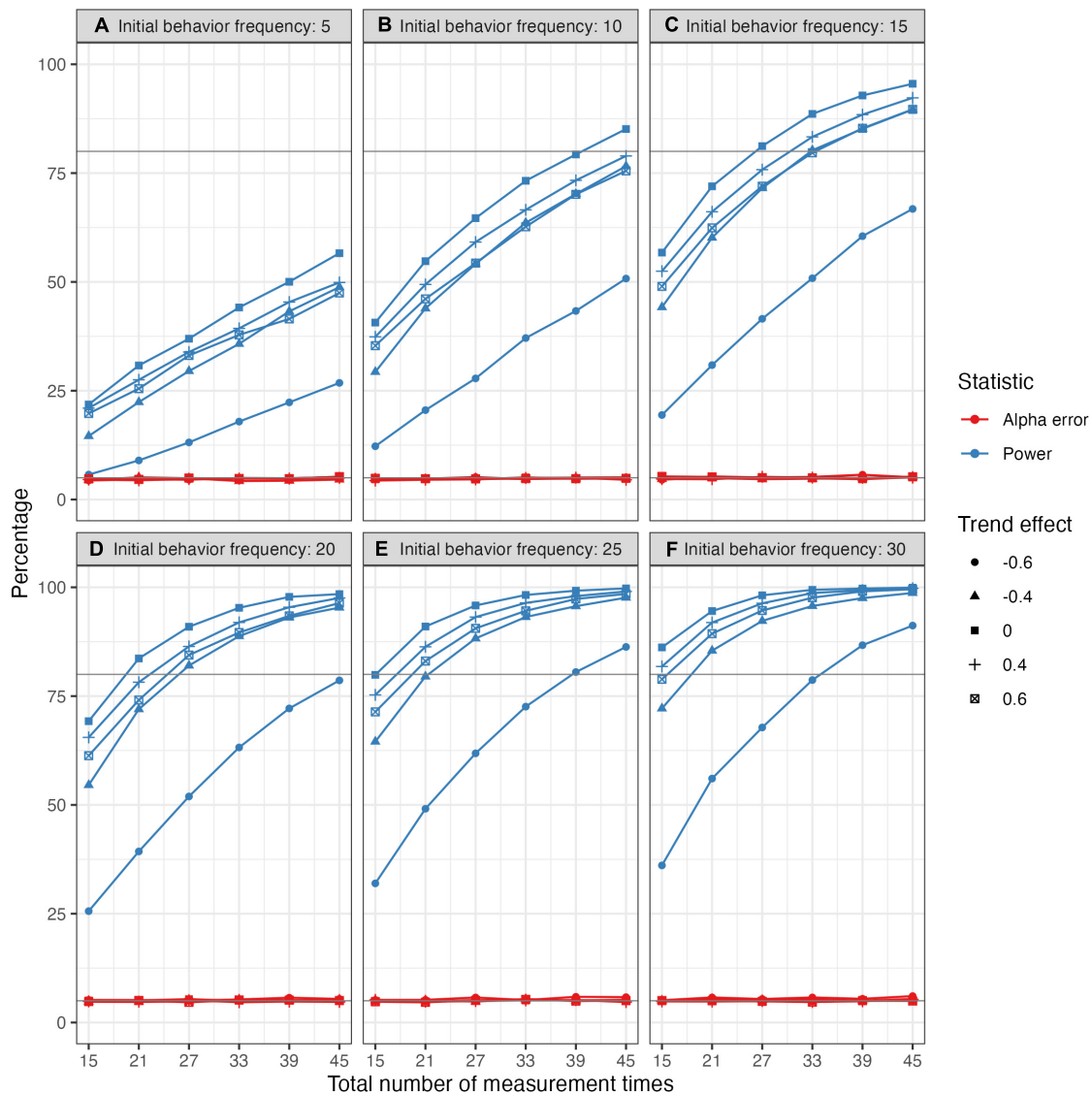


FIGURE 5 | Power and alpha error probability (line colour) for trend effects (dot shape), initial behavior frequency (part), and number of measurement times (x-axis) (simulation 3).

guide researchers on crucial elements of SCEDs to prevent unfavorable decisions.

In our study, the level effect of the intervention had a powerful influence on statistical power. Strong effects, where the behavior was reduced by 60% or higher, were correctly identified under almost all conditions. However, the exact characteristics played a crucial role when the intervention effects were medium or low. Effects that were equivalent to a reduction of 20% could not be correctly identified (independent of the design characteristics). Prior knowledge about the intervention's expected effect size might help researchers make research design decisions that lead to higher statistical power. However, such knowledge might not be available for all kinds of interventions. Moreover, the expected intervention effect is not something researchers have control over. Therefore, the following discussion will primarily focus on

parameters that are at least under partial control of the researcher, designing and conducting the study.

Initial Behavior Frequency

In contrast to the effect size, researchers *can* influence the operationalization of the outcome variable. A dependent variable can be operationalized differently, leading to different outcome variable frequencies (e.g., a higher sampling rate or larger observation intervals for each measurement time). This is an asset for researchers as the results of our study indicate that the outcome variable frequency has a substantial impact on statistical power, too. Low initial behavior frequencies set high demands on the number of measurement times required to correctly identify effects (especially when the intervention effect is small) or might even wholly prevent its identification (initial behavior

frequency ≤ 5). Based on our results, we would recommend targeting operationalizations that allow initial problem behavior frequencies greater than 20. Such frequencies are in line with the existing state of research in SCED (Pustejovsky et al., 2019).

Number of Measurement Times in Phase A

The number of measurement times is one of the main design elements of SCEDs that the investigator can influence. Our study indicates that the length of phase A has a significant influence on the resulting test power: Low numbers of measurement times in phase A (≤ 7), which are common, hinder the identification of even strong intervention effects (60% reduction). Nonetheless, such low numbers of measurement times (e.g., 3) depict the lower end of the recommendations in the relevant literature (e.g., Kratochwill et al., 2013). This suggests that many published single-case studies have low power due to a short phase A length. It is better to prolong phase A than phase B in those cases. This seems to be a particularly relevant finding, as researchers might feel forced to begin an intervention (phase B) as quickly as possible due to ethical (stressful classroom situation) or economic (costs which come along with the extension of phase A) reasons. However, our results emphasize the need to extend phase A (even under challenging conditions) as the costs for a short phase A might be the failure to identify a potentially helpful intervention. Extending phase B cannot compensate for a low number of measurement times in phase A. Based on our results, we recommend at least nine measurement times during phase A when the estimated intervention effect is an estimated reduction of 60% or more. When the reduction is between 40% and 60%, collect data for at least 15 measurement times in phase A and extend phase B to at least 30 measurement times.

Number of Measurement Times in Phase B

A similar pattern of results occurs when focusing on the number of measurement times in phase B. Again, an increment in the number of measurement times leads to an overall increase in statistical power. However, the number of measurement times in phase A and intervention effect size seem to be of higher relevance (given a reasonable number of at least 15 measurements in phase B). This implies that extending phase B does not improve statistical power to a sufficient level if the number of measurement times in phase A is too small. For smaller intervention effects (i.e., a reduction of 40%), the length of phase B seems of additional relevance when the length of phase A increases.

Trend Effect

Depending on the situation, one can make assumptions about the presence, intensity, and direction of a data trend (e.g., when researchers receive information about the student's development prior to the study). In many situations, however, trend effects are difficult to predict. Our results suggest that piecewise Poisson-regressions are robust to the possible influence of trend effects (i.e., the results showed no increased alpha error risk even

when very strong trend effects were prevalent). Nevertheless, a strong negative trend effect (i.e., a reduction of 60% across all measurements of a single case) affects test power. Since this finding occurs mainly in situations where the initial frequency of the behavior is low, a possible explanation could be a floor effect (e.g., due to the data trend frequencies being so low that the intervention effect cannot develop its full strength). Since trend effects thus might play an important role in predicting test power, it seems crucial to control for the presence of such effects during data analysis. Here, the results of a piecewise regression analysis might help detect a strong trend effect after the data collection. Recognizing a data trend could subsequently serve as further evidence for a potential limitation of test power.

The results of our study clearly emphasize the power of piecewise Poisson-regressions in analyzing SCED studies. Despite the usefulness of the chosen analytical approach, it becomes clear that important design specifications must be considered. Despite our efforts to derive some guiding principles, it becomes clear that the test power depends on an intricate interplay between various design specifications. What an adequate single-case experimental design looks like depends on the context, the type of intervention, and the behavior to address. As with all other hypothesis-testing research designs, researchers planning SCED studies should include power analyses in their research planning. Factors such as the number of measurement times or the precise operationalization of the dependent variable can often be adjusted to improve the design of studies from the very beginning. In addition, *post hoc* power analyses also help to provide at least a rough estimate of the statistical power and uncover the strength and caveats of a design. Based on our results, it additionally becomes clear that the characteristics of SCEDs that come along with high test power deviate from common practice, especially regarding the number of measurement times.

Limitations

Despite the insights gained, the study at hand has some limitations. First, our insights are limited to a specific scenario (i.e., count data; an intervention aiming at a frequency reduction), which cannot be generalized to all potential scenarios that might occur in practice. Therefore, additional simulation studies addressing other scenarios are recommended. Specifically, our intervention effect only comprised a level effect and no additional slope effect. However, a slope effect might occur (depending on the interventional approach). Second, we focused on AB designs as the essential ingredient of many SCED variants. In research practice, AB designs only represent one design among other SCEDs. Therefore, the validity of our results is restricted to AB designs.

Implications for Analysis of Single-Case Experimental Designs

We focused on the use of regression analysis in this study. Other procedures exist to estimate phase differences in SCED data, such as overlap indices or randomization tests. Our results are not simply generalizable to these procedures. However, we would argue that the power of these procedures is no higher than that

of the regression analyses analyzed here. Thus, the requirements for achieving sufficient power are likely to be even higher. Fortunately, software packages are available today to calculate exact power estimations for specific design specifications. All analyses in this study have been calculated with the R package *scan* (Wilbert and Lüke, 2022), which also allows for calculating the power for different SCEDs (e.g., multiple baseline and multiphase designs; gaussian or binomial distributed data) and other methods of data analysis (e.g., randomization tests or Tau-U).

Based on the result of our analyses, we would like to recommend that researchers conduct *a priori* power analysis for any SCED they are planning. If the intended research design yields insufficient power (usually below 80%) or the alpha-error probability is too high (usually above 5%), two optional modifications to the SCED can increase the power of the design: (1) Increasing the number of measurement times, especially in phase A (often phase A is too short). (2) Implement a more sensitive operationalization that increases the frequency of the dependent variable (ideally to an initial frequency of at least 20). In addition, conducting a multiple-baseline design with three or more cases/situations or adding a second A and B phase (withdrawal design) may also increase the statistical power of the design.

Researchers cannot and will not always optimize decisions regarding their specific research design in favor of statistical power. Sometimes, the specific circumstances in which SCEDs are applied prevent this (e.g., ethical reasons, opportunities to implement an intervention in the institutional context). Whenever possible, however, we consider it necessary for

research in SCEDs to take into account the test power and alpha-error probability and, accordingly, to conduct only those studies that can realistically detect an existing intervention effect. We believe that it would be beneficial in the future to present and demand considerations of statistical power for publications reporting SCEDs as well.

DATA AVAILABILITY STATEMENT

All data and analyses presented in this manuscript are publicly available in the Open Science Framework: <https://osf.io/ys3a9/> and <https://files.eric.ed.gov/fulltext/ED510743.pdf>.

AUTHOR CONTRIBUTIONS

JW did the conceptualization, carried out the data curation, formal analysis, and software, investigated the data, performed the methodology, visualized the data, wrote the original draft, and wrote, reviewed, and edited the manuscript. MB-R and TL did the conceptualization, investigated the data, performed the methodology, wrote the original draft, and wrote, reviewed, and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number: 491466077.

REFERENCES

- Allison, D. B. (1992). When cyclicity is a concern: a caveat regarding phase change criteria in single-case designs. *Compr. Ment. Health Care* 2, 131–149.
- Beretvas, S. N., and Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: methodological issues and practice. *Evid. Based Commun. Assess. Interv.* 2, 129–141. doi: 10.1080/17489530802446302
- Briesch, A. M., and Briesch, J. M. (2016). Meta-analysis of behavioral self-management interventions in single-case research. *School Psychol. Rev.* 45, 3–18. doi: 10.17105/spr45-1.3-18
- Busacca, M. L., Anderson, A., and Moore, D. W. (2015). Self-management for primary school students demonstrating problem behavior in regular classrooms: evidence review of single-case design research. *J. Behav. Educ.* 24, 373–401. doi: 10.1007/s10864-015-9230-3
- Davis, D. H., Gagné, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., and Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behav. Modif.* 37, 62–89. doi: 10.1177/0145445512453734
- Dowdy, A., Hantula, D. A., Travers, J. C., and Tincani, M. (2022). Meta-analytic methods to detect publication bias in behavior science research. *Perspect. Behav. Sci.* 45, 37–52. doi: 10.1007/s40614-021-00303-0
- Dugard, P., File, P., and Todman, J. (2012). *Single-Case and Small-n Experimental Designs: A Practical Guide to Randomization Tests*, 2nd Edn. New York, NY: Routledge.
- Edgington, E., and Onghena, P. (2007). *Randomization Tests*, 4th Edn. Boca Raton, FL: CRC Press.
- Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behav. Res. Methods Instrum. Comput.* 34, 324–331. doi: 10.3758/BF03195459
- Greenwald, A. G. (1976). Within-subjects designs: to use or not to use? *Psychol. Bull.* 83, 314–320.
- Harrison, J. R., Soares, D. A., Rudzinski, S., and Johnson, R. (2019). Attention deficit hyperactivity disorders and classroom-based interventions: evidence-based status, effectiveness, and moderators of effects in single-case design research. *Rev. Educ. Res.* 89, 569–611. doi: 10.3102/0034654319857038
- Heyvaert, M., and Onghena, P. (2014). Randomization tests for single-case experiments: state of the art, state of the science, and state of the application. *J. Contextual Behav. Sci.* 3, 51–64. doi: 10.1016/j.jcbs.2013.10.002
- Hintze, J. M., Volpe, R. J., and Shapiro, E. S. (2002). Direct Observation of Student Behavior. *Best Pract. School Psychol.* 4, 993–1006.
- Hitchcock, J. H., Horner, R. H., Kratochwill, T. R., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2014). The what works clearinghouse single-case design pilot standards: who will guard the guards? *Remedial Spec. Educ.* 35, 145–152.
- Huitema, B. E. (1986). “Statistical analysis and single-subject designs,” in *Research Methods in Applied Behavior Analysis: Issues and Advances*, eds A. Poling and R. W. Fuqua (New York, NY: Plenum), 209–232.
- Huitema, B. E., and McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educ. Psychol. Meas.* 60, 38–58. doi: 10.1177/00131640021970358
- Jones, R. R., Weinrott, M. R., and Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *J. Appl. Behav. Anal.* 11, 277–283.
- Keppel, G. (1982). *Design and Analysis: A Researcher's Handbook*, 2nd Edn. Englewood Cliffs, NJ: Prentice Hall.
- Klapproth, F. (2018). Biased predictions of students' future achievement: an experimental study on pre-service teachers' interpretation of curriculum-based measurement graphs. *Stud. Educ. Eval.* 59, 67–75. doi: 10.1016/j.stueduc.2018.03.004

- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2010). *Single-Case Designs Technical Documentation*.
- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2013). Single-case intervention research design standards. *Remedial Spec. Educ.* 34, 26–38. doi: 10.1177/0741932512452794
- Lane, J. D., and Ledford, J. R. (2014). Using interval-based systems to measure behavior in early childhood special education and early intervention. *Top. Early Child. Special Educ.* 34, 83–93. doi: 10.1177/0271121414524063
- Ledford, J. R., Lane, J. D., and Gast, D. L. (2018). *Dependent Variables, Measurement, and Reliability: Single Case Research Methodology*. New York, NY: Routledge.
- Maggin, D. M., Briesch, A. M., and Chafouleas, S. M. (2013). An application of the what works clearinghouse standards for evaluating single-subject research: synthesis of the self-management literature base. *Remedial Spec. Educ.* 34, 44–58. doi: 10.1177/0741932511435176
- Matyas, T. A., and Greenwood, K. M. (1990). Visual analysis of single-case time series: effects of variability, serial dependence, and magnitude of intervention effects. *J. Appl. Behav. Anal.* 23, 341–351. doi: 10.1901/jaba.1990.23-341
- Moeyaert, M., Ferron, J. M., Beretvas, S. N., and Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *J. Sch. Psychol.* 52, 191–211. doi: 10.1016/j.jsp.2013.11.003
- Moeyaert, M., Klingbeil, D. A., Rodabaugh, E., and Turan, M. (2021). Three-level meta-analysis of single-case data regarding the effects of peer tutoring on academic and social-behavioral outcomes for at-risk students and students with disabilities. *Remedial Spec. Educ.* 42, 94–106. doi: 10.1177/0741932519855079
- Nock, M. K., Michel, B. D., Photos, V. I., and McKay, D. (2007). "Single-case research designs," in *Handbook of Research Methods in Abnormal and Clinical Psychology*, ed. D. McKay (Thousand Oaks, CA: Sage Publications), 337–350.
- Parker, R. I., and Brossart, D. F. (2003). Evaluating single-case research data: a comparison of seven statistical methods. *Behav. Ther.* 34, 189–203.
- Parker, R. I., and Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *J. Behav. Educ.* 21, 254–265. doi: 10.1007/s10864-012-9153-1
- Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *J. School Psychol.* 68, 99–112. doi: 10.1016/j.jsp.2018.02.003
- Pustejovsky, J. E., Swan, D. M., and English, K. W. (2019). An examination of measurement procedures and characteristics of baseline outcome data in single-case research. *Behav. Modif.* [Epub ahead of print]. doi: 10.1177/0145445519864264
- Shadish, W. R., Hedges, L. V., Horner, R. H., and Odom, S. L. (2015). *The Role of Between-Case Effect Size in Conducting, Interpreting, and Summarizing Single-Case Research*. Washington, DC: National Center for Education Research.
- Shadish, W. R., Hedges, L. V., and Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: a primer and applications. *J. Sch. Psychol.* 52, 123–147. doi: 10.1016/j.jsp.2013.11.005
- Shadish, W. R., Kyse, E. N., and Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: new applications and some agenda items for future research. *Psychol. Methods* 18, 385–405. doi: 10.1037/a0032964
- Shadish, W. R., and Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behav. Res. Methods* 43, 971–980. doi: 10.3758/s13428-011-0111-y
- Smith, J. D. (2012). Single-case experimental designs: a systematic review of published research and current standards. *Psychol. Methods* 17, 510–550. doi: 10.1037/a0029312
- Soares, D. A., Harrison, J. R., Vannest, K. J., and McClelland, S. S. (2016). Effect size for token economy use in contemporary classroom settings: a meta-analysis of single-case research. *Sch. Psychol. Rev.* 45, 379–399. doi: 10.17105/spr45-4.379-399
- Travers, J. C., Cook, B. G., Therrien, W. J., and Coyne, M. D. (2016). Replication research and special education. *Remedial Spec. Educ.* 37, 195–204. doi: 10.1177/0741932516648462
- Vannest, K. J., and Sallese, M. R. (2021). Benchmarking effect sizes in single-case experimental designs. *Evid. Based Commun. Assess. Interv.* 15, 142–165. doi: 10.1080/17489539.2021.1886412
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.
- Wilbert, J. (2022). *scplot: An R Package for Visualizing Single-Case Data*. Potsdam: University of Potsdam.
- Wilbert, J., Bosch, J., and Lüke, T. (2021). Validity and judgement bias in visual analysis of single-case data. *Int. J. Res. Learn. Disabil.* 5, 13–24. doi: 10.28987/ijrld.5.1.13
- Wilbert, J., and Lüke, T. (2022). *Scan: Single-Case Data Analyses for Single and Multiple Baseline Designs*. Potsdam: University of Potsdam.
- Wolfe, K., Barton, E. E., and Meadan, H. (2019). Systematic protocols for the visual analysis of single-case research data. *Behav. Anal. Pract.* 12, 491–502. doi: 10.1007/s40617-019-00336-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Wilbert, Börnert-Ringleb and Lüke. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

EDITED BY

Sarah Powell,
University of Texas at Austin,
United States

REVIEWED BY

Stefan Blumenthal,
University of Rostock, Germany
Jana Jungjohann,
University of Regensburg, Germany

*CORRESPONDENCE

Tatjana Leidig
tleidig@uni-koeln.de

SPECIALTY SECTION

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

RECEIVED 10 April 2022

ACCEPTED 28 June 2022

PUBLISHED 29 July 2022

CITATION

Leidig T, Casale G, Wilbert J,
Hennemann T, Volpe RJ, Briesch A and
Grosche M (2022) Individual,
generalized, and moderated effects
of the good behavior game on at-risk
primary school students: A multilevel
multiple baseline study using
behavioral progress monitoring.
Front. Educ. 7:917138.
doi: 10.3389/feduc.2022.917138

COPYRIGHT

© 2022 Leidig, Casale, Wilbert,
Hennemann, Volpe, Briesch and
Grosche. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Individual, generalized, and moderated effects of the good behavior game on at-risk primary school students: A multilevel multiple baseline study using behavioral progress monitoring

Tatjana Leidig ^{1*}, Gino Casale ², Jürgen Wilbert ³,
Thomas Hennemann ¹, Robert J. Volpe ⁴,
Amy Briesch ⁴ and Michael Grosche ²

¹Department of Special Education and Rehabilitation, Faculty of Human Sciences, University of Cologne, Cologne, Germany, ²Institute of Educational Research, University of Wuppertal, Wuppertal, Germany, ³Institute of Inclusive Education, Faculty of Human Sciences, University of Potsdam, Potsdam, Germany, ⁴Department of Applied Psychology, Northeastern University, Boston, MA, United States

The current study examined the impact of the Good Behavior Game (GBG) on the academic engagement (AE) and disruptive behavior (DB) of at-risk students' in a German inclusive primary school sample using behavioral progress monitoring. A multiple baseline design across participants was employed to evaluate the effects of the GBG on 35 primary school students in seven classrooms from grade 1 to 3 ($M_{\text{age}} = 8.01$ years, $SD_{\text{age}} = 0.81$ years). The implementation of the GBG was randomly staggered by 2 weeks across classrooms. Teacher-completed Direct Behavior Rating (DBR) was applied to measure AE and DB. We used piecewise regression and a multilevel extension to estimate the individual case-specific treatment effects as well as the generalized effects across cases. Piecewise regressions for each case showed significant immediate treatment effects for the majority of participants (82.86%) for one or both outcome measures. The multilevel approach revealed that the GBG improved at-risk students' classroom behaviors generally with a significant immediate treatment effect across cases (for AE, $B = 0.74$, $p < 0.001$; for DB, $B = -1.29$, $p < 0.001$). The moderation between intervention effectiveness and teacher ratings of students' risks for externalizing psychosocial problems was significant for DB ($B = -0.07$, $p = 0.047$) but not for AE. Findings are consistent with previous studies indicating that the GBG is an appropriate classroom-based intervention for at-risk students and expand the literature regarding differential effects for affected students. In addition, the study supports

the relevance of behavioral progress monitoring and data-based decision-making in inclusive schools in order to evaluate the effectiveness of the GBG and, if necessary, to modify the intervention for individual students or the whole group.

KEYWORDS

classroom behavior, good behavior game, multilevel analysis, piecewise regression, single case design

Introduction

The national prevalence rates of mental health problems in Germany indicate that approximately 17 to 20% of schoolchildren demonstrate psychosocial problems (i.e., externalizing problems such as aggressive behavior or hyperactivity, and internalizing problems such as depressiveness or anxiety) with various degrees of severity (Barkmann and Schulte-Markwort, 2012; Klipker et al., 2018). These problems can negatively affect classroom behaviors and the social and academic performance of individual students (Kauffman and Landrum, 2012) and their peers (Barth et al., 2004). As the majority of students with behavioral problems in Germany are educated in inclusive schools without special education services (Volpe et al., 2018), teachers need strategies to successfully deal with their behavior problems, promote social-emotional development, and continuously evaluate the effectiveness of the implemented strategies.

Research reviews have demonstrated that group contingencies such as the Good Behavior Game (GBG; Barrish et al., 1969) are effective interventions for managing externalizing behavior problems in general education classrooms (Maggin et al., 2017; Fabiano and Pyle, 2019). The GBG is an easy-to-use interdependent group contingency intervention with extensive empirical support that utilizes students' mutual dependence to reduce problem behaviors and to improve prosocial and academic behaviors (Flower et al., 2014). The main features of this game are easily comprehensible (Flower et al., 2014): (1) selecting goals and rules, (2) recording rule violations, (3) explaining the rules of the game and determining the rewards, (4) dividing the students in two or more teams to play against each other, and (5) playing the GBG for a specified amount of time. In the classic variant, the team receives a mark ("foul") when a team member breaks a rule. At the end of the game, the team with the fewest fouls wins. In multi-tiered-systems of support (MTSS; Batsche, 2014), the GBG is typically integrated in regular classroom instruction in tier 1 as a part of a proactive classroom management approach (Simonsen and Myers, 2015). When the GBG is played in conjunction with behavioral progress monitoring, this combination provides an appropriate way to facilitate data-based decision-making in supporting students at risks for emotional and behavioral disorders (EBD) within

a MTSS: If an individual student does not respond to the GBG, this information can be used to decide whether further interventions should be added in tier 2 (Donaldson et al., 2017). In Germany, evidence-based practice within MTSS is still in its infancy (e.g., Voß et al., 2016; Hanisch et al., 2019); this also applies to globally known interventions such as the GBG and data-based decision-making based on behavioral progress monitoring. Given the potential benefits of the GBG for students with or at risk for EBD, additional information is needed on its impact and suitability in a German population.

Evidence base of the good behavior game

For group design studies in general, the GBG meta-analysis by Flower et al. (2014) indicated moderate effects (Cohen's $d = 0.50$) on problem behaviors (e.g., aggression, off-task behavior, and talking out), whereas the meta-analysis of randomized-controlled trials of the GBG by Smith et al. (2021) resulted in only small but significant effect sizes for conduct problems (i.e., aggression or oppositional behavior) (Hedges' $g = 0.10$, $p = 0.026$) and moderate but not significant effect sizes for inattention (i.e., concentration problems and off-task behavior) (Hedges' $g = 0.49$, $p = 0.123$). For conduct problems, the comparatively smaller effects are likely due to the rigorous standards of randomized-controlled trials, implying, for instance, less biased estimates of study effects; for inattention, this may reflect the significant heterogeneity of the findings and the overall small number of included studies (Smith et al., 2021). Meta-analyses of single-case research of the GBG (Bowman-Perrott et al., 2016; Flower et al., 2014) and class-wide interventions for supporting student behavior (Chaffee et al., 2017) found a significant and immediate treatment effect for reducing challenging behaviors (e.g., disruptive behavior, aggression, off-task behavior, talking out, and out-of-seat) (Flower et al., 2014: $\beta = -0.2038$, $p < 0.01$) and medium to high effects across both general and special education settings (Bowman-Perrott et al., 2016: $\text{TauU} = 0.82$; Chaffee et al., 2017: $\text{TauU} = 1.00$) with larger effects on disruptive and off-task behavior (e.g., out-of-seat, talking out, interrupting, pushing and fighting) ($\text{TauU} = 0.81$) than on-task behavior (e.g., working quietly, following teacher's instructions, and getting

materials without talking) ($\text{Tau}U = 0.59$), and higher effects for students with or at risk for EBD ($\text{Tau}U = 0.98$) than for students without any difficulties ($\text{Tau}U = 0.76$) (Bowman-Perrott et al., 2016). Although the meta-analyses revealed the effectiveness of the GBG across settings, it must be noted that the majority of research examined the impact of the GBG in general education settings with typically developing students (Moore et al., 2022). Even though the studies included often lack concrete information on the implementation of mainstreaming or inclusive education in the sample, against the background of the development of the school systems since the 1990s, it can be assumed that many studies of the last 30 years were conducted in mainstreaming or inclusive settings.

When interpreting the meta-analytic findings, some methodological aspects have to be considered. In some studies using a group design, students were nested within classrooms. This approach is helpful to infer the effectiveness for a population of students based on inferential statistics. However, it does not automatically allow to distinguish which students the GBG was effective for and which students did not benefit at all (or even increased their problem behavior). In fact, in group research designs, the behavior of individual students differing from the mean of all students is considered a measurement error, and the groups' effect size does not tell us anything about case-specific treatment effects (Lobo et al., 2017). In principle, this also applies to subgroup analyses, even if they examine the effects more specifically. Furthermore, as Smith et al. (2021) noted, the available group design studies and single case studies measure similar but finally different outcomes: While the results of group design studies allow conclusions regarding general, cross-situational changes in student behavior (trait), no conclusions can be drawn about the effects of GBG on targeted behavior in the classroom situations in which GBG is played (state). The reverse is equally true. Against this background, the comparability of results of group studies and single-case studies is constrained. In addition, regarding the meta-analyses of single-case research, it is important to consider that most of the included studies investigated the impact of the GBG on a group or a class, not on individual students (Donaldson et al., 2017).

Evidence base for effects on at-risk students in single-case research using progress monitoring

There is limited information about the impact of the GBG on students with or at risk for EBD in single-case research using progress monitoring. As Bowman-Perrott et al. (2016) critically noted, most studies included in their meta-analysis focused on typically developing students with a normal range of disruptive behaviors. From a methodological perspective, the information provided by single-case research is limited due to the composition and nature of the samples and the methods of data analysis used. With only a few exceptions, the vast majority of single-case studies investigated a class or a group as a single-case and analyzed data at the classroom level (Donaldson et al., 2017). This approach is helpful to test the positive impact of

the GBG on a group as a whole, but it does not allow analysis at the level of individual students (Donaldson et al., 2017; Foley et al., 2019). Furthermore, the sample sizes are relatively small (i.e., three to 12 cases), and the studies varied widely with regard to sample characteristics (e.g., setting, classroom size, class composition). As a result, neither the findings of individual studies nor the aforementioned meta-analyses can simply be generalized to specific populations of students, e.g., at-risk students (Bowman-Perrott et al., 2016; Donaldson et al., 2017).

In only three of the four studies reporting individual student data included in the meta-analysis by Bowman-Perrott et al. (2016), the target students were explicitly identified as the most challenging students in their class (i.e., students displaying more disruptive behaviors than peers) (Medland and Stachnik, 1972; Tanol et al., 2010; Hunt, 2012). We found six further studies investigating the impact of the GBG on students with or at risk for EBD at the level of individual students (Donaldson et al., 2017; Groves and Austin, 2017; Pennington and McComas, 2017; Wiskow et al., 2018; Foley et al., 2019; Moore et al., 2022). In seven of the aforementioned nine studies, all of the target students showed improvements, although to differing degrees. However, Donaldson et al. (2017) and Hunt (2012) found individual non-responders. Furthermore, Donaldson et al. (2017) and Moore et al. (2022) reported decreasing positive effects of the GBG over time for some children who frequently exhibited disruptive behavior. As Donaldson et al. (2017) concluded, only teachers who play the GBG with progress monitoring at the level of individual students can avoid unnecessary implementation of individualized interventions, and identify students who need additional support beyond the class-wide intervention.

In the few studies that have examined the effects of the GBG at the level of individual students, one can analyze the differential impact on individual students, but one cannot conclude whether the GBG is effective for a specific group of students, e.g., with or at risk for EBD. No study at the individual level or at the group level used an inferential statistic approach to examine whether the effects are statistically significant, which is important to generalize the results (Shadish et al., 2013). Even in single-case research, researchers are not only interested in specific individual treatment effects to support evidence-based decisions, but also whether the effects can be generalized to other cases. To our knowledge, no study evaluating the efficacy of the GBG has used inferential statistics to generalize the average treatment effect across cases within the same study. Accordingly, neither the results of studies at classroom level nor those at the level of individual students are representative for students with or at risk for EBD, so that the evidence for this group remains comparatively weak (Bowman-Perrott et al., 2016; Donaldson et al., 2017). In addition to case-specific inferential statistics, multilevel models enable investigating the overall effectiveness across participants considering the average

treatment effect, variations across cases and possible influential factors (Moeyart et al., 2014).

Psychosocial problems as moderator for the effects of the good behavior game

Additional research on the students' individual characteristics is needed to assess the possible influential factors (Bowman-Perrott et al., 2016; Maggin et al., 2017). The results of group design studies suggest that students' psychosocial problems (i.e., externalizing problems such as aggressive behavior or hyperactivity, and internalizing problems such as depressiveness or anxiety) are associated with the effectiveness of the GBG: They indicate differential effects of the GBG on externalizing and internalizing problems depending on the students' individual risk levels and types (e.g., van Lier et al., 2005; Kellam et al., 2008; Spilt et al., 2013) with partly contradictory results for students with or at risk for severe externalizing behavior problems (e.g., aggressive, violent, and criminal behavior) and for students with combinations of risks (e.g., combination of social and behavior risks). Analyzing single-case research, Bowman-Perrott et al. (2016) identified the EBD risk status as a potential moderator for the effect on externalizing classroom behaviors (i.e., larger effects for students with EBD or at risk for EBD), however without further investigation of the risk level and subtype of behavior problems. Therefore, the role of students' psychosocial problems as a potential moderator requires further investigation.

The current study

In sum, the existing research adds different pieces to the puzzle whether GBG is an evidence-based intervention, partly also for students with or at risk for EBD (Bowman-Perrott et al., 2016). There is some support for the effectiveness of the GBG either for individual students or for groups of students, e.g., with or at risk for EBD (Joslyn et al., 2019; Smith et al., 2021). However, our study is the first study about the GBG that uses a large sample of individual at-risk students nested in classrooms and tested the hypothesis of individual, general, and differential effectiveness within the same study using inferential statistics on individual and group level at the same time. Typically, the number of cases in single-case research varies between 1 and 13 (Shadish and Sullivan, 2011). Our large sample of 35 students and our methodological approach give us the opportunity to examine case-specific treatment effects as well as effects across cases under stable conditions. Combining group statistics and single-case statistics makes all Council of Exceptional Children's standards for classifying evidence-based special education interventions (Cook et al., 2015) applicable for our study (see [Supplementary Material](#)).

Therefore, the current study is designed to extend previous research by investigating the impact of the GBG on at-risk students' classroom behaviors in inclusive settings in Germany

using behavioral progress monitoring and analyzing the data using regression analyses to estimate case-specific treatment effects as well as effects across students. In particular, we investigate the interaction between the impact of the GBG and students' psychosocial problems due to the potential influence of students' individual risks. We hypothesized moderate to large effects on at-risk students' academic engagement and disruptive behavior for the majority of students and across all cases with an immediate treatment effect, and only a small additional slope effect. Furthermore, we anticipate that students' psychosocial problems moderate the intervention effects. We hypothesize that the impact of the GBG would interact with the magnitude of students' behavioral problems. Specifically, we assume that the higher students' externalizing problems are, the more effective the GBG will be.

Method

Participants and setting

In Germany, inclusion in schools is primarily understood as learning together of students with and without special educational needs in general school settings (KMK, 2011). This study was conducted in an inclusive primary school in a midsize town in western Germany (North Rhine-Westphalia) with 12 first-through-fourth-grade classrooms. Due to the legal requirements of North Rhine-Westphalia, support in the areas of learning, speech and behavior can be provided in all grades regardless of the formally identified special educational needs, so that there is no administrative ascriptive diagnosis for some of the students with special educational needs. Furthermore, as a rule, no formal diagnosis of special educational needs in learning, behavior, and speech is provided for first and second graders. In the inclusive primary school in our study, approximately 50 children were with or at risk for learning disabilities and/or EBD. In line with official school records, approximately 70% of all students had a migration background.

After the introduction of the project during the teachers' conference, seven general education teachers decided to participate in the study with their classes. Each of these classroom teachers nominated five students based on their professional experience and judgment as the most challenging students in their classroom ($n = 35$; five first graders, 15 second graders and 15 third graders; five girls and 30 boys). Based on teacher ratings on the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), 22 (63%) of the nominated students showed externalizing risks, whereas only two (6%) demonstrated internalizing risks, two students (6%) showed risks in both areas, and nine students (26%) exposed no psychosocial risks. Further demographic information and the SDQ data are summarized in [Table 1](#).

After consultation with the teachers to identify the classroom situation believed to be the most problematic

TABLE 1 Student demographic information and data of the behavioral screening instrument (SDQ).

Classroom (Grade)	Student	Gender	Age	Migration background	SDQ	
					Externalizing score ^a	Internalizing score ^b
A (1 st)	S1	Boy	6	0	9 ⁺	1
	S2	Boy	6	0	11 ⁺⁺	0
	S3	Girl	7	0	6	0
	S4	Boy	7	1	5	3
	S5	Boy	6	1	7 ⁺	1
B (2 nd)	S6	Boy	7	1	8 ⁺	1
	S7	Boy	8	1	4	3
	S8	Boy	7	1	13 ⁺⁺⁺	3
	S9	Boy	7	1	15 ⁺⁺⁺	6 ⁺
	S10	Boy	7	0	18 ⁺⁺⁺	4
C (2 nd)	S11	Girl	7	1	10 ⁺⁺	3
	S12	Boy	7	0	13 ⁺⁺⁺	1
	S13	Boy	9	0	13 ⁺⁺⁺	3
	S14	Boy	7	1	10 ⁺⁺	4
	S15	Boy	8	1	8 ⁺	1
D (2 nd)	S16	Boy	7	1	4	0
	S17	Boy	7	0	6	3
	S18	Boy	7	0	9 ⁺	4
	S19	Boy	7	1	4	7 ⁺
	S20	Girl	7	1	6	10 ⁺⁺⁺
E (3 rd)	S21	Boy	9	1	3	1
	S22	Boy	8	1	6	0
	S23	Boy	9	0	8 ⁺	5
	S24	Boy	9	0	6	1
	S25	Boy	8	0	12 ⁺⁺⁺	3
F (3 rd)	S26	Girl	9	0	13 ⁺⁺⁺	2
	S27	Boy	8	1	11 ⁺⁺	2
	S28	Girl	8	1	7 ⁺	3
	S29	Boy	8	1	10 ⁺⁺	1
	S30	Boy	8	0	10 ⁺⁺	9 ⁺⁺
G (3 rd)	S31	Boy	8	0	6	2
	S32	Boy	8	1	9 ⁺	0
	S33	Boy	8	0	9 ⁺	3
	S34	Boy	9	1	11 ⁺⁺	1
	S35	Boy	8	0	10 ⁺⁺	0

For the migration background, 0 = no and 1 = yes. For SDQ externalizing and internalizing scores, ⁺ = slightly raised, ⁺⁺ = high and ⁺⁺⁺ = very high.

^aCategorization by Goodman et al. (2010): 7-9 = slightly raised, 10-11 = high, 12-20 = very high. ^bCategorization by Goodman et al. (2010): 6-7 = slightly raised, 8-9 = high, 10-20 = very high.

concerning externalizing classroom behaviors, we defined the target instructional period for playing the GBG for each classroom (i.e., individual work in math for classroom B, D, and G, and individual work in German language for classrooms A, C, E, and F).

Measures

Direct behavior rating

Direct Behavior Rating (DBR; Christ et al., 2009) was used to measure *academic engagement* (AE) and *disruptive behavior* (DB). DBR has already been adapted, evaluated, and

implemented for German classrooms using the operational definitions of AE and DB provided by Chafouleas (2011) (Casale et al., 2015, 2017). AE included students' active or passive participation in ongoing academic activities such as engaging appropriately in classroom activities, concentrated working, completing tasks on time and raising their hands. DB was defined as behaviors disrupting others or affecting students' own or other students' learning, such as speaking without permission, leaving one's seat, noisemaking, having undesirable private discussions and fooling around. For both dependent variables, we used a single-item scale (SIS) with these broadly defined items representing a common behavior class (Christ et al., 2009). Previous

research focusing on the German DBR scales supports their generalizability and dependability across different raters (i.e., general classroom teachers and special education teachers), items, and occasions (Casale et al., 2015, 2017). These studies showed that the DBR-SIS provides dependable scores ($\Phi > 0.70$) for an individual student's behavior after four measurement occasions. Furthermore, measurement invariance testing across high-frequency occasions with short intervals showed the sensitivity of DBR (Gebhardt et al., 2019). In addition, previous research from the United States supports the reliability, validity and sensitivity of DBR to changes to the SIS regarding the targeted behaviors (e.g., Briesch et al., 2010; Chafouleas et al., 2012). The teachers observed the behavior of the five nominated students in the target classroom situation during baseline and intervention. Immediately after each observation, the teachers rated both AE and DB for each of the five students on a paper-pencil questionnaire using a 6-point Likert-type scale (i.e., 0 = *never*, 5 = *always*).

Strengths and difficulties questionnaire for teachers

Participants' internalizing and externalizing problems were assessed with the German version of the worldwide used behavioral screening questionnaire SDQ (Goodman, 1997) for teachers. It consists of 25 items equally divided across five subscales (hyperactivity, conduct problems, emotional problems, peer problems, and prosocial behavior). Besides the 5-factor model, a 3-factor model containing the factors of externalizing problems, internalizing problems, and prosocial behavior is used to interpret results (Goodman et al., 2010). Evaluations of the German version indicate an acceptable fit both of the 5-factor model (Bettge et al., 2002) and the 3-factor model (DeVries et al., 2017), and a good internal consistency for the teacher version of the SDQ (Cronbach's α between 0.77 and 0.86; Saile, 2007). After each item was scored by teachers on a 3-point Likert-type scale (0 = *not*

true, 1 = *somewhat true*, 3 = *certainly true*), we calculated the internalizing and externalizing subscales by summing the conduct and hyperactivity scales for the externalizing subscale and summing the emotional and peer problems scales for the internalizing subscale (Goodman et al., 2010).

Design and procedures

A multiple baseline across-classroom design with a 2-week staggered randomized-phase start was used. Each classroom was randomly assigned to a single *a priori* designated intervention start point (Kratochwill and Levin, 2010). After the baseline phase, which was previously determined to be at least 8 but not more than 38 days, the intervention phase (70 to 100 days) started with a 2-week interval between the groups depending on the start point. Due to holidays and cancelled lessons because of school events, full-day conferences, part-time employment of teachers, and in-service training, the number of days with the opportunity to play the GBG fluctuated between 61 to 87 days. The GBG was played *de facto* 3 to 5 days a week. Phase lengths for each classroom are displayed in Table 2.

Teacher training and mentoring

To support an effective implementation, we developed a program with the components shown in Figure 1 according to recommendations from the literature (Hagermoser Sanetti et al., 2014; Poduska and Kurki, 2014) and considering the specific school setting. In addition, the teachers received training on the use of DBR based on Chafouleas (2011), consisting of the theoretical background, the practical use of DBR, and feedback by a research assistant.

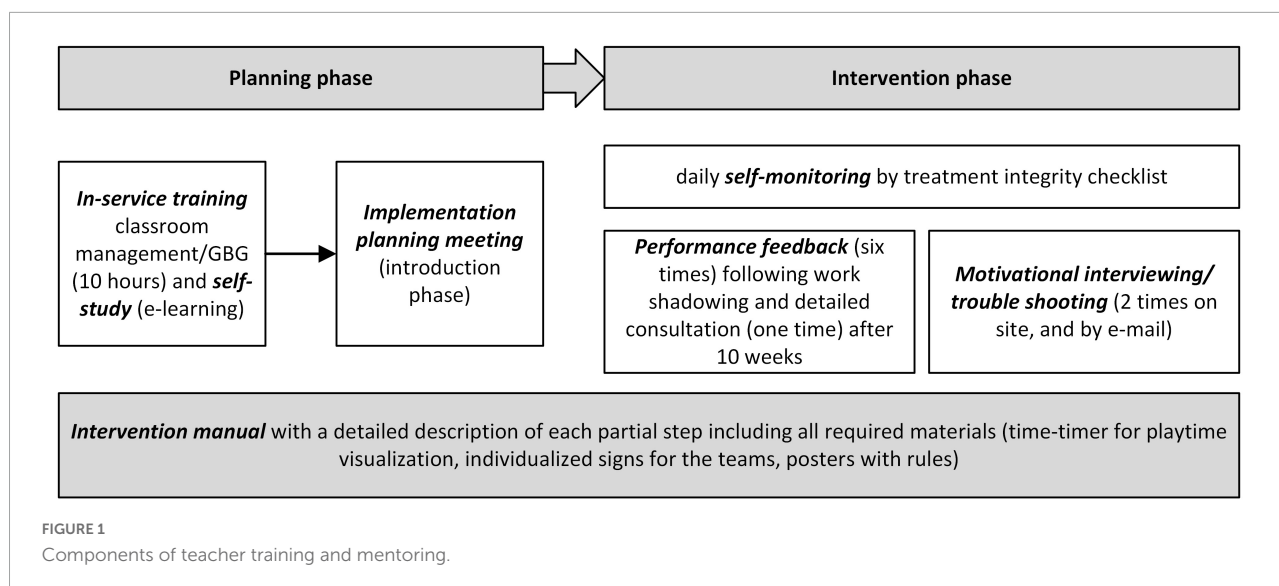
In total, the training before the implementation of the GBG included 14 hours for classroom management and GBG (i.e., 10 hours in-service training and 4 hours self-study) and 5 hours for DBR (in-service training). Self-study was implemented as an e-learning course with text, audio, video, and

TABLE 2 Design data, treatment integrity, and treatment usability for each classroom.

Classroom	A	B	Treatment Intensity ^a	Treatment Integrity ^b	Treatment Usability (M)			
					Acceptability	Understanding	Collaboration	Feasibility
A	8	100	97.70%	92.86%	5.29	4.75	5.33	5.17
B	8	100	79.76%	100%	5.43	5.25	6.00	5.17
C	18	90	87.01%	100%	4.29	4.75	4.67	4.67
D	38	70	81.97%	89.29%	4.86	4.75	5.33	5.00
E	18	90	70.89%	100%	4.71	5.25	5.00	4.67
F	28	80	83.10%	88.10%	4.71	4.50	5.67	4.67
G	28	80	64.79%	88.57%	5.00	5.00	5.00	4.83

A = baseline phase length in days; B = intervention phase length in days; M = mean values. For treatment usability, higher scores are indicative of more extreme responding in the direction of the scale assessed.

^aTreatment intensity as number of days the GBG was played based on the adjusted opportunities to play, expressed as percentage. ^bTreatment integrity rated by observer.



interactive elements. The in-service training featured input with video examples, discussion, and small group practice including feedback from both trainers and peers. In addition, the teachers received a detailed printed intervention manual with all steps and materials. Immediately before the start of the intervention phase, details regarding the introduction phase of the GBG and specific individual questions were clarified in a two-hour implementation planning meeting. During the intervention phase, we provided six brief appointments (15 to 20 minutes) for performance feedback spread over the entire intervention period following work shadowing, in each case combined with talks about pending issues and difficulties, and one appointment (1 hour) for a detailed consultation after ten weeks based on the results of the behavioral progress monitoring. Motivational interviewing/trouble shooting was offered two times on site by the trainers (1 hour) and additionally by email.

Baseline

Teachers completed the paper-pencil-based DBR-SIS for each student each day at the end of the target instructional period. The assessment units were 10 minutes long.

Intervention

The classic version of the GBG was played (Flower et al., 2014) with dividing the students in two or more teams to play against each other, marking “fouls” for rule-breaking, and rewarding the team with the fewest fouls as winner of the game. The teachers divided their class into five to six teams using the existing assigned group seating arrangements, with five to six students on each team. The teams were maintained throughout the intervention phase. Contrary to previous practices reported by Donaldson et al. (2017), no student was placed on his or her own team because of behavior, e.g. disruptive behavior. If

problems arose on a team, the teacher spoke with the team to find common solutions. After the teams had given themselves names (e.g., “Lions” and “The cool kids”), each team received a sign with the team name, which was affixed to the team table. A poster at the front of the classroom listed three to four rules that had been previously jointly compiled (e.g., “I am quiet” – “I work intently” – “I sit in my seat”). Moreover, a timer was placed near the poster to be clearly visible to all teams, and the names of teams were written on the board to mark the fouls. Respecting the specific classroom situation and the relationship with the students, the goals, rules, rule violations, and rewards slightly deviated across classrooms (e.g., rewarding with goodies such as chocolate or gummy bears, small prizes such as stickers or pens, or activities such as group games or extra reading time). After 10 minutes playing in the defined period, the teachers counted the fouls, named the winning team and delivered the reward to the winning team.

Treatment integrity

At six measurement points, trained observers collected treatment integrity data on the accuracy with which the teacher implemented the GBG using a seven-item treatment integrity checklist adapted from the Treatment Integrity Planning Protocol (TIPP; Hagermoser Sanetti and Kratochwill, 2009) with dichotomous ratings (*agreement – disagreement*). In addition, the teachers completed an analogous treatment integrity checklist with additional options for comments (e.g., reasons for not playing the game or information on problems) and noted the number of fouls overall and per team on each day. The total adherence across all classrooms throughout the six measurement points was 94.12% (range = 88.1% to 100%). For these days, the interobserver agreement (observer – teacher) was 100%. Teachers’ ratings over the whole treatment period

showed that the overall treatment integrity measured by daily self-assessment was 96.15% (range = 85.5% to 100%).

Treatment usability

Treatment usability from the teachers' points of view was assessed with the German version of the Usage Rating Profile (URP; Briesch et al., 2017) consisting of 20 items loading on four factors. The teachers indicated the extent to which they agreed with each of the items using a 6-point Likert scale (i.e., 1 = *strong disagreement*, 6 = *strong agreement*). The results of the URP revealed agreement regarding the usage of the GBG intervention across all teachers for each factor: $M = 4.9$ ($SD = 0.38$) for acceptability, $M = 4.9$ ($SD = 0.28$) for understanding, $M = 5.2$ ($SD = 0.45$) for home-school collaboration, and $M = 4.88$ ($SD = 0.23$) for feasibility. Treatment integrity and usability are shown in Table 2.

Data analysis

To examine the impact of the GBG on AE and DB on each single case and across cases, descriptive statistics and inferential statistics were used to analyze the data. Therefore, after calculating phase means and effect sizes for each single case, we conducted regression analyses for each single case and across cases to replicate the results as well as to obtain statistical significance and overall quantification (Manolov and Moeyaert, 2017). The multilevel approach to estimate the overall effects was also used to analyze the interaction between students' psychosocial problems and the intervention effects. First, we reported descriptive statistics and calculated the non-rescaled non-overlap of all pairs (NAP; see Alresheed et al., 2013) with medium effects indicated by values of 66% to 92%, and strong effects indicated by values of 93% to 100%. Subsequently, we analyzed the data using a piecewise regression approach (Huitema and McKean, 2000). This procedure enables the control of developmental trends in the data (trend effects) and the differentiation between continuous (slope effect) and immediate (level effects) intervention effects. We conducted piecewise regressions for each single case and a multilevel extension (see Van den Noortgate and Onghena, 2003; Moeyart et al., 2014) for all cases with measurements at level 1 nested in subjects at level 2. The multilevel analyses were set up as a random intercept and random slope model with all three parameters (trend, slope, and level effects) as fixed and random factors. To test the significance of the random effects, we applied a likelihood ratio test for each random slope factor comparing the full model against a model without the target factor. To analyze the moderating effects of students' internalizing and externalizing problems on the intervention effects, we inserted cross-level interactions into the model. All analyses were conducted using R (R Core Team, 2018) and the scan package (Wilbert and Lüke, 2018).

Results

Descriptive statistics

Table 3 summarizes descriptive statistics, including the NAP as a common effect size for both AE and DB. The non-rescaled NAP indicated a medium or strong effect for both dependent variables for all participants, varying between 77.8% and 99.8% for AE and between 78.3% and 100% for DB. In detail, there were strong effects for 12 (34.29%) and medium effects for 23 (65.71%) participants regarding AE, and strong effects for 18 (51.54%) and medium effects for 17 (48.46%) participants regarding DB.

Inferential statistics

We analyzed the data case by case, conducting piecewise regression analyses to calculate the impact of the GBG on AE and DB for each case. Subsequently, we used a multilevel extension to calculate the impact of the GBG for all cases with measurements at level 1 nested in subjects at level 2.

Piecewise regression for each single case

Fourteen cases showed both significant level increases in AE and decreases in DB ($p < 0.05$). Three participants (S2, S3, and S20) showed significant level increases only for AE ($p < 0.05$), whereas 12 participants demonstrated significant level decreases only for DB ($p < 0.05$). The slope effect was significant for increases in AE as well as decreases in DB for three participants ($p < 0.05$ for S10, S26, and S33). Four participants (S19, S21, S29, and S35) demonstrated neither a significant level nor a significant slope effect for one of the dependent variables ($p > 0.05$). The results for each case are shown in Table 4.

Multilevel analyses

First, we conducted multilevel analyses of all single cases. Overall, for AE and DB, we found significant level effects. On average, AE increased by 0.74 points ($p < 0.001$) on a 6-point Likert-type scale, and DB decreased by 1.29 points ($p < 0.001$). Furthermore, the significant slope effect for DB indicated a decrease of 0.01 points ($p = 0.002$) per measurement occasion. The similar increasing slope effect for AE failed to reach statistical significance ($p = 0.055$). Second, we calculated the random effects regarding the variability between cases. We found significant level effects for both of the dependent variables. For AE, the estimated standard deviation for the level effect was $SD = 0.65$ ($p < 0.001$), whereas for DB, it was $SD = 0.81$ ($p < 0.001$). The estimated standard deviation for the slope effect was $SD = 0.00$ ($p = 1.000$) for AE and $SD = 0.15$ ($p = 0.034$) for DB. The results for both fixed and random effects are shown in Table 5.

TABLE 3 Descriptive statistics for the 35 single-cases for both dependent variables.

Case	n_A	n_B	Academic Engagement					Disruptive Behavior				
			mis_A	mis_B	M_A (SD)	M_B (SD)	NAP (%)	mis_A	mis_B	M_A (SD)	M_B (SD)	NAP (%)
S1	8	100	1	24	2.0 (1.2)	3.5 (1.1)	88.6	1	24	1.1 (0.4)	0.2 (0.4)	93.9
S2	8	100	1	17	2.1 (1.2)	3.6 (1.1)	86.9	1	18	1.1 (0.7)	0.4 (0.8)	83.8
S3	8	100	1	20	3.6 (0.8)	4.1 (0.8)	77.8	1	20	1.1 (0.7)	0.2 (0.6)	90.2
S4	8	100	1	20	3.4 (0.5)	4.7 (0.7)	95.9	1	21	2.3 (1.1)	0.1 (0.3)	99.4
S5	8	100	1	19	3.1 (1.1)	4.3 (0.8)	87.8	1	20	1.1 (0.9)	0.1 (0.3)	88.7
S6	8	100	1	36	1.9 (0.9)	3.5 (0.8)	94.2	1	38	2.1 (1.1)	0.0 (0.2)	99.8
S7	8	100	2	41	2.3 (1.2)	2.5 (1.2)	79.6	2	41	1.8 (0.8)	0.2 (0.4)	98.6
S8	8	100	1	38	1.9 (1.3)	3.2 (0.9)	88.9	1	38	2.1 (1.1)	0.0 (0.1)	99.9
S9	8	100	1	34	1.4 (0.8)	3.6 (0.8)	97.8	1	35	3.7 (0.5)	0.0 (0.3)	100.0
S10	8	100	2	41	2.2 (1.5)	3.3 (0.9)	88.5	2	42	1.0 (0.6)	0.0 (0.1)	96.1
S11	18	90	5	33	1.9 (0.3)	2.6 (0.6)	90.0	5	34	2.9 (0.9)	0.2 (0.7)	98.7
S12	18	90	5	43	3.8 (0.6)	4.4 (0.7)	89.5	5	43	2.7 (0.9)	0.2 (0.5)	99.3
S13	18	90	7	37	2.4 (0.5)	2.7 (0.5)	88.3	7	37	1.8 (0.6)	0.2 (0.5)	98.7
S14	18	90	5	31	2.4 (0.7)	3.0 (0.5)	88.1	5	32	1.8 (0.7)	0.1 (0.3)	99.4
S15	18	90	5	34	2.9 (0.3)	3.7 (0.5)	93.7	5	35	1.8 (0.7)	0.1 (0.4)	98.9
S16	38	70	11	23	3.7 (0.5)	4.1 (0.6)	84.5	11	23	1.5 (0.8)	0.8 (0.7)	86.9
S17	38	70	11	20	3.7 (0.7)	4.3 (0.8)	86.2	11	22	1.0 (0.7)	0.4 (0.5)	87.0
S18	38	70	13	28	4.0 (0.5)	4.7 (0.5)	91.1	13	29	0.7 (0.6)	0.1 (0.4)	90.6
S19	38	70	13	26	3.1 (0.8)	4.1 (0.8)	91.7	13	27	0.6 (0.8)	0.3 (0.5)	85.3
S20	38	70	11	20	3.5 (0.7)	4.7 (0.6)	94.2	11	22	0.1 (0.3)	0.0 (0.0)	78.3
S21	18	90	7	38	3.7 (0.6)	3.5 (0.7)	79.9	7	40	0.9 (0.7)	0.7 (0.8)	85.4
S22	18	90	4	36	3.6 (0.7)	4.4 (0.6)	88.9	5	38	1.4 (0.9)	0.6 (0.6)	90.3
S23	18	90	4	37	3.3 (0.6)	3.7 (0.9)	85.0	4	38	1.3 (0.6)	0.5 (0.6)	90.5
S24	18	90	4	43	4.0 (0.8)	4.4 (0.6)	84.9	4	45	1.3 (0.6)	0.6 (0.6)	90.7
S25	18	90	3	55	3.3 (0.8)	4.2 (0.6)	93.6	3	57	2.0 (0.7)	0.7 (0.6)	97.3
S26	28	80	12	29	3.3 (0.8)	3.3 (0.6)	82.2	12	29	2.0 (1.1)	1.7 (0.8)	84.5
S27	28	80	12	26	3.5 (0.6)	3.4 (0.7)	79.3	12	26	2.1 (1.0)	1.8 (0.9)	83.6
S28	28	80	13	29	2.1 (0.5)	2.7 (0.8)	90.4	13	29	3.1 (0.7)	2.0 (0.9)	94.0
S29	28	80	14	32	2.4 (0.6)	3.0 (0.9)	91.2	14	32	2.8 (0.8)	1.6 (0.9)	94.8
S30	28	80	16	30	3.2 (0.6)	3.9 (0.7)	93.6	16	30	2.1 (0.9)	0.6 (0.7)	96.7
S31	28	80	12	36	2.7 (1.3)	4.8 (0.6)	97.7	12	36	1.2 (1.4)	0.1 (0.4)	92.4
S32	28	80	12	36	3.2 (1.2)	4.9 (0.3)	97.4	12	36	1.1 (0.7)	0.0 (0.2)	96.5
S33	28	80	12	35	1.9 (1.1)	4.9 (0.4)	99.8	12	36	1.7 (1.2)	0.1 (0.3)	96.2
S34	28	80	12	36	4.2 (0.6)	4.9 (0.4)	93.8	12	36	0.6 (0.6)	0.0 (0.2)	92.5
S35	28	80	12	36	4.1 (0.8)	4.8 (0.6)	93.1	12	36	0.4 (0.6)	0.1 (0.5)	88.6

n = number of data points, which was the same for all of the students within the same group in each classroom; mis = missing data points; M = mean, SD = standard deviation; NAP = non-rescaled non-overlap of all pairs. Missing data points resulted from the school and teacher factors (e.g., public and movable holidays, cancelled lessons, part-time employment, and illness) as well as student illness, incomplete DBR, etc.

Interaction with students' psychosocial problems

We inserted cross-level interactions between level effects and the different SDQ subscales into the model. For AE, we did not find significant interactions with the level of the intervention phase for either the SDQ total score ($B = 0.03$, $p = 0.292$) or one of the subscales ($B_{int} = 0.03$, $p = 0.505$; $B_{ext} = 0.03$, $p = 0.423$). The analysis for DB showed a significant interaction for the externalizing subscale ($B = -0.07$, $p = 0.047$) and no significant interactions for the other scales ($B_{total} = -0.05$, $p = 0.076$;

$B_{int} = -0.02$, $p = 0.927$). Table 6 contains the results for both dependent variables.

Discussion

The purpose of the current study was to evaluate the impact of the GBG on at-risk students' AE and DB in an inclusive primary school in Germany using behavioral progress monitoring. Extending previous studies, we used

TABLE 4 Piecewise regression analyses for the 35 single cases for both dependent variables.

Case	Academic Engagement			Disruptive Behavior		
	Trend	Level	Slope	Trend	Level	Slope
S1	-0.22	2.12*	0.22	0.08	-1.31*	-0.08
S2	-0.23	1.85*	0.24	0.08	-0.71	-0.09
S3	-0.08	1.26*	0.07	-0.16	-0.65	0.16
S4	-0.09	1.58*	0.09	0.12	-2.68*	-0.12
S5	0.03	0.82	-0.03	-0.3*	-0.06	0.3*
S6	0.23	0.72	-0.23	-0.23*	-1.34*	0.23*
S7	-0.15	0.14	0.16	-0.06	-1.5*	0.06
S8	-0.06	0.84	0.07	-0.13*	-1.64*	0.13*
S9	-0.16	2.4*	0.17	-0.09*	-3.33*	0.09*
S10	-0.36*	1.93*	0.36*	0.11*	-1.33*	-0.11*
S11	0.03	1.16*	-0.04	-0.07	-2.3*	0.07
S12	0.06	-0.59	-0.04	-0.05	-2.08*	0.05
S13	0	0.36	0	-0.04	-1.14*	0.04
S14	0.02	0.62*	-0.02	0.07*	-2.06*	-0.07*
S15	0.03	0.34	-0.02	-0.02	-1.45*	0.02
S16	-0.01	0.56*	0.01	0.02	-1.21*	-0.01
S17	-0.01	0.54	0.02	0	-0.73*	0.01
S18	0	0.53*	0	-0.01	-0.55*	0.01
S19	0.03*	-0.18	-0.02	-0.02	0.04	0.02
S20	0.01	0.74*	-0.01	0	-0.16	0
S21	-0.05	0.06	0.05	0.05	-1.03	-0.04
S22	-0.06	1.28*	0.05	0.11*	-1.94*	-0.11*
S23	0	0.11	0.01	0.02	-0.83*	-0.02
S24	-0.05	0.77*	0.05	0.04	-1.17*	-0.03
S25	-0.01	0.62	0.01	0.03	-0.99*	-0.04
S26	-0.05*	0.39	0.05*	0.07*	-1.4*	-0.06*
S27	-0.01	-0.15	0.01	-0.06*	0.36	0.06*
S28	-0.01	0.51	0.01	0	-1.21*	0
S29	-0.02	0.08	0.04	-0.01	-0.27	-0.01
S30	-0.03	0.52	0.04	0.02	-1.44*	-0.02
S31	0.01	1.88*	-0.01	0.05*	-1.77*	-0.05*
S32	-0.01	1.7*	0.01	0.01	-1.02*	-0.01
S33	-0.04*	3.51*	0.04*	0.06*	-2.38*	-0.06*
S34	-0.01	0.7*	0.02	0	-0.48*	0
S35	0	0.55	0	-0.03	0.1	0.02

* = significant ($p < 0.05$).

piecewise regression for each of the 35 single cases and a multilevel extension to examine both level and slope effects. Furthermore, we examined the interaction with students' psychosocial problems as potential influencing factors moderating the effectiveness.

Main findings

The individual-level data analyses revealed that the majority of at-risk students benefited from the GBG. Whereas the non-rescaled NAP indicated a medium or strong effect for both dependent variables for all participants, the inferential statistics did not reveal statistically significant improvements for all cases. Piecewise regressions for each single case enabled us to identify

significant immediate treatment effects for 14 participants for both outcomes (40.0%), for three participants for AE only (8.57%), and for 12 participants for DB only (34.29%). However, six students (17.14%) showed no significant level effects for AE or DB. The results at the individual level support and extend the limited prior research, indicating that the classroom-based GBG intervention is effective for improving at-risk students' classroom behaviors, but there are students who do not respond to the intervention (Hunt, 2012; Donaldson et al., 2017; Moore et al., 2022). Consistent with previous research reporting classroom-wide data (Flower et al., 2014; Bowman-Perrott et al., 2016), the multilevel approach revealed that the GBG improved AE and reduced DB for students with challenging behavior with a significant immediate treatment effect across cases. In line with the meta-analysis by Bowman-Perrott et al. (2016), the GBG was more effective in reducing DB than increasing AE. Similar to Flower et al. (2014), we found statistically significant slightly decreasing DB throughout the intervention phase. In contrast to the single case studies by Donaldson et al. (2017) and Moore et al. (2022), in which some participants with or at risk for EBD showed slightly increasing trends over the course of the intervention, this slope effect indicates a continuous and decreasing change in DB.

The findings from the present study extend prior meta-analysis results (Bowman-Perrott et al., 2016) by investigating students' psychosocial risks as potential moderators for the impact of the intervention. We found significant moderating effects for students' externalizing problems on the intervention effect for DB, meaning that the higher students' externalizing problems were, the more effective the GBG was. These findings correspond with the results of reviews on the subject of interventions for aggressive behavior (Waschbusch et al., 2019) as well as longitudinal studies evidencing the GBG as an effective

TABLE 5 Fixed and random effects of the multilevel piecewise-regression models for academic engagement and disruptive behavior.

Parameter	Fixed effects				Random effects		
	B	SE	t	p	Estimated SD	L	p
Academic Engagement							
Intercept	3.02	0.14	21.83	< 0.001**	0.74	109.59	< 0.001**
Trend	0.00	0.00	-0.92	0.358	0.01	1.94	0.746
Level Phase B	0.74	0.13	5.86	< 0.001**	0.65	65.71	< 0.001**
Slope Phase B	0.01	0.00	1.92	0.055	0.00	0.02	1.000
Disruptive Behavior							
Intercept	1.50	0.13	11.63	< 0.001**	0.71	210.92	< 0.001**
Trend	0.01	0.00	3.19	0.001**	0.02	12.90	0.012**
Level Phase B	-1.29	0.15	-8.77	< 0.001**	0.81	138.84	< 0.001**
Slope Phase B	-0.01	0.00	-3.18	0.002**	0.15	10.43	0.034*

* = significant ($p < 0.05$); ** = significant ($p < 0.001$).

TABLE 6 Interaction of psychosocial problems with the impact of the good behavior game.

Scale	<i>B</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	<i>B</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Academic Engagement						Disruptive Behavior				
SDQ total										
Intercept	2.94	0.12	2376	24.25	< 0.001**	1.61	0.12	2349	12.24	< 0.001**
Trend	0.00	0.00	2376	3.72	< 0.001**	0.00	0.00	2349	0.30	0.766
Level Phase B	0.67	0.17	2376	5.30	< 0.001**	−1.20	0.13	2349	−9.33	< 0.001**
SDQ total	−0.06	0.03	33	−2.22	0.034	0.04	0.03	33	1.40	0.172
Level Phase B*SDQ total	0.03	0.03	2376	1.06	0.292	−0.05	0.03	2349	−1.77	0.076
SDQ int										
Intercept	2.94	0.13	2376	22.99	< 0.001**	1.61	0.14	2349	11.76	< 0.001**
Trend	0.00	0.00	2376	3.70	< 0.001**	0.00	0.00	2349	0.34	0.738
Level Phase B	0.67	0.13	2376	5.32	< 0.001**	−1.20	0.05	2349	−9.15	< 0.001**
SDQ int	−0.05	0.05	33	−0.94	0.356	−0.01	0.05	33	−0.09	0.927
Level Phase B*SDQ int	0.03	0.05	2376	0.67	0.505	−0.02	0.05	2349	−0.39	0.697
SDQ ext										
Intercept	2.94	0.12	2376	23.88	< 0.001**	1.61	0.13	2349	12.60	< 0.001**
Trend	0.00	0.00	2376	3.70	< 0.001**	0.00	0.00	2349	0.30	0.766
Level Phase B	0.67	0.13	2376	5.37	< 0.001**	−1.20	0.13	2349	−9.41	< 0.001**
SDQ ext	−0.07	0.04	33	−1.98	0.056	0.07	0.04	33	1.93	0.062
Level Phase B*SDQ ext	0.03	0.03	2376	0.80	0.423	−0.07	0.04	2349	−1.99	0.047*

SDQ total = SDQ total score; SDQ int = SDQ subscale internalizing problems; SDQ ext = SDQ subscale externalizing problems; * = significant ($p < 0.05$); ** = significant ($p < 0.001$).

intervention to reduce aggressive behavior for children with high risks in general (van Lier et al., 2005) and in boys with persistent high risks (Kellam et al., 2008). However, even though the majority of students with externalizing risks benefited from the intervention, three of the non-responders showed high risks assessed by their teachers. This finding leads to the assumption that for children with high externalizing risks, further individual factors moderate the effectiveness of group contingencies (Maggin et al., 2017). No effects were found for internalizing risks. Considering that the majority of the students nominated by teachers showed no internalizing problems, we cannot deduce effects for students with internalizing risks from our study.

Overall, our findings indicated that the impact of the classroom-based GBG program varied as a function of individual children. For both outcomes, and in particular for AE, there was large variability between individuals. There are several explanations for this finding. Although no clear pattern emerges in our data, it is possible that aspects of treatment integrity and usability may have affected the outcomes, particularly among the non-responders (Moore et al., 2022). From a methodological point of view, for some students in our sample, the high AE values at baseline led to minimal room for improvement, suggesting a possible ceiling effect (Ho and Yu, 2015). Likewise, there might have been floor effects for some of the students with low DB at baseline. Furthermore, it is important to consider the situation in which the GBG was played. It is possible that the target situation was not the most difficult part of the lesson for all of the nominated students; thus, their AE and, in part, their DB might not have been as problematic as usual. In addition,

struggling with learning strategies, as a common problem of students with challenging behavior (Kauffman and Landrum, 2012), could affect the effectiveness of the GBG regarding AE. The implementation of additional components such as self-monitoring strategies (Bruhn et al., 2015) could be necessary to increase AE for non-responding students (Smith et al., 2021).

Interestingly, we could not find any interaction of externalizing problems and intervention effectiveness for AE. In addition to the aforementioned possible ceiling effects, the similarity of the measured constructs must be considered. Although hyperactivity and conduct problems can negatively impact school functioning and academic performance (Mundy et al., 2017), the externalizing subscales of the SDQ as a standardized measure for assessing child mental health problems are more closely linked to disruptive classroom behaviors than to academic engagement as a typical school-related construct. Further associated factors, such as psychological and cognitive dimensions of engagement, including having sense of belonging or motivational beliefs (Wang and Eccles, 2013) or the level of academic enabling skills (Fabiano and Pyle, 2019), could moderate the intervention effects on AE.

Limitations

The findings from this study should be interpreted by considering several potential limitations. First, not all of the nominated students in our sample had high ratings in the externalizing and/or internalizing subscales of the SDQ.

Therefore, it can be assumed that not all of the nominated students were at risk for EBD. Although research has indicated that teachers are competent in identifying students in their classrooms with problem behaviors (Lane and Menzies, 2005), we believe that in addition to teacher nomination, future research should use other methods to identify students with challenging behavior (i.e., systematic behavioral assessment in the baseline).

Second, teachers both delivered the intervention and rated students' performance. The 'double burden' of teaching and rating as well as the teachers' acceptance of the intervention could affect their ratings. However, despite limited associations between behavioral change and acceptability, research has demonstrated the sensitivity of the DBR-SIS completed by implementing teachers (Chafouleas et al., 2012; Smith et al., 2018). Furthermore, we were unable to conduct systematic direct observations by trained observers or video recordings. As such, we decided to use the DBR-SIS as an efficient tool with acceptable reliability, validity and sensitivity within our aims.

Third, our dependent variables were broad categories combining different behaviors. These target behaviors enabled us to compare our results, particularly with the findings of the existing meta-analysis. On the other hand, students' individual changes in specific behaviors could not be tested. Furthermore, we only investigated one potential moderator for the impact of the intervention. However, the differentiated analysis of these risks for two key aspects of classroom behaviors extends previous research independently of the need for further investigation.

Implications for research and practice

Analyzing the responses to the GBG using behavioral progress monitoring helps to identify students who need additional support on tier 2 in a MTSS and to recognize at an early stage if positive effects are decreasing over time for individual students (Donaldson et al., 2017; Moore et al., 2022). The selected combination of data analysis methods enables precise alignment with our aims and considers the characteristics of the data (Manolov and Moeyaert, 2017): Due to our large sample and the analytical method chosen, we were able to investigate the impact of the GBG on individual at-risk student classroom behaviors as well as its effectiveness across cases. We therefore believe that our study is a methodologically sound study about the GBG that could be used to substantiate the evidence classification of the GBG. Considering the fact that the single-case studies of the GBG work with small samples and use methods for data analysis that do not address the problem of autocorrelation, misestimates of effectiveness due to the methods chosen are plausible (Shadish et al., 2014), and the number of non-responders tends to be underestimated. The number of non-responders indicates

that we should be very careful with the transfer of results of studies investigating a class or a group as a single-case to the behavioral development of individual students with challenging behavior. Therefore, further single-case research with larger-than-usual samples and meta-analytical approaches are necessary to extend the findings regarding the impact of the GBG on at-risk students' classroom behaviors as well as further potential moderators. As shown in our study, externalizing risks seem to moderate the impact of the GBG on reducing disruptive behavior. Thus, our findings imply that in future intervention studies, the effects should be controlled for possible influences of externalizing risks. Furthermore, in addition to potential individual factors, functional characteristics (Maggin et al., 2017), environmental moderators, such as the classroom level of aggression (Waschbusch et al., 2019), peer factors, and school climate (Farrell et al., 2013), should be examined.

Overall, the results of our study suggest that the GBG facilitates at-risk students' behavioral development in inclusive settings in Germany. In particular, in our sample, students who were assessed by teachers as exhibiting high externalizing behavior problems can benefit. These results should encourage teachers to implement this classroom-based intervention and to monitor its effect using behavioral progress monitoring. Our findings also suggest that a combination of the behavioristic method of the GBG with cognitive methods such as self-monitoring (Bruhn et al., 2015) could be necessary to enable the effects regarding AE. In our study, treatment integrity was high, and teachers assessed the intervention as suitable for their daily work. However, whether the intervention can be sustainably implemented depends on several factors. Despite the simple rules of the game, coaching throughout the implementation and positive impacts, teachers do not see sufficient possibilities to integrate the GBG naturally in their daily work (Coombes et al., 2016). Furthermore, the teachers in our study likewise reported that they did not find the time to play the GBG due to school events or learning projects. In addition, some found it difficult to consistently implement behavioral progress monitoring. To increase sustainability, the maintenance of both the GBG and progress monitoring should be carefully planned and monitored. If this is successful, the GBG, in conjunction with behavioral progress monitoring, is an appropriate classroom-based intervention to improve at-risk students' classroom behaviors and to adjust students' supports data-based at an early stage.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval were not required in accordance with the local legislation and institutional requirements. Following the school law and the requirements of the ministry of education of the federal state North Rhine-Westphalia (Schulgesetz für das Land Nordrhein-Westfalen), school administrators decided in co-ordination with their teachers about participation in this scientific study. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. Verbal informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

MG, GC, RV, JW, and TL: methodology. JW: software. JW and TL: formal analysis. GC and TL: investigation. TH, GC, and TL: resources. TL, GC, and JW: data curation. TL: writing – original draft and visualization. TH and RV: supervision. GC and TL: project administration. All authors: conceptualization and writing – review and editing.

Funding

This study was supported by the Marbach Residency Program (Jacobs Foundation). We acknowledge support for the

Article Processing Charge from the DFG (German Research Foundation, 491454339).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2022.917138/full#supplementary-material>

References

- Alresheed, F., Hott, B. L., and Bano, C. (2013). Single Subject Research: a Synthesis of Analytic Methods. *JOSEA* 2, 1–18.
- Barkmann, C., and Schulte-Markwort, M. (2012). Prevalence of Emotional and Behavioural Disorders in German Children and Adolescents: a Meta-Analysis. *J. Epidemiol. Community Health* 66, 194–203. doi: 10.1136/jech.2009.102467
- Barrish, H. H., Saunders, M., and Wolf, M. M. (1969). Good Behavior Game: Effects of Individual Contingencies for Group Consequences on Disruptive Behavior in a Classroom. *J. Appl. Behav. Anal.* 2, 119–124. doi: 10.1901/jaba.1969.2-119
- Barth, J. M., Dunlap, S. T., Dane, H., Lochman, J. E., and Wells, K. C. (2004). Classroom Environment Influences on Aggression, Peer Relations, and Academic Focus. *J. Sch. Psychol.* 42, 115–133. doi: 10.1016/j.jsp.2003.11.004
- Batsche, G. (2014). "Multi-Tiered Systems of Supports for Inclusive Schools," in *Handbook of Effective Inclusive Schools*, eds J. McLeskey, N. L. Waldron, F. Spooner, and B. Algozzine (New York, NY: Routledge), 183–196.
- Bettge, S., Ravens-Sieberger, U., Wietzker, A., and Hölling, H. (2002). Ein Methodenvergleich der Child Behavior Checklist und des Strengths and Difficulties Questionnaire. [Comparison of Methods Between the Child Behavior Checklist and the Strengths and Difficulties Questionnaire]. *Gesundheitswesen* 64, 119–124. doi: 10.1055/s-2002-39264
- Bowman-Perrott, L., Burke, M. D., Zaini, S., Zhang, N., and Vannest, K. (2016). Promoting Positive Behavior Using the Good Behavior Game: a Meta-Analysis of Single-Case Research. *J. Posit. Behav. Interv.* 18, 180–190. doi: 10.1177/1098300715592355
- Briesch, A. M., Casale, G., Grosche, M., Volpe, R. J., and Hennemann, T. (2017). Initial Validation of the Usage Rating Profile-Assessment for Use Within German Language Schools. *Learn. Disab.* 15, 193–207.
- Briesch, A. M., Chafouleas, S. M., and Riley-Tillman, T. C. (2010). Generalizability and Dependability of Behavior Assessment Methods to Estimate Academic Engagement: a Comparison of Systematic Direct Observation and Direct Behavior Rating. *School Psych. Rev.* 39, 408–421.
- Bruhn, A., McDaniel, S., and Kreigh, C. (2015). Self-Monitoring Interventions for Students With Behavior Problems: a Systematic Review of Current Research. *Behav. Disord.* 40, 102–121. doi: 10.17988/BD-13-45.1
- Casale, G., Grosche, M., Volpe, R. J., and Hennemann, T. (2017). Zuverlässigkeit von Verhaltensverlaufsdiagnostik über Rater und Messzeitpunkte bei Schülern mit externalisierenden Verhaltensproblemen. [Dependability of Direct Behavior Ratings Across Rater and Occasion in Students with Externalizing Behavior Problems]. *Empirische Sonderpädagogik* 9, 143–164.
- Casale, G., Hennemann, T., Volpe, R. J., Briesch, A. M., and Grosche, M. (2015). Generalisierbarkeit und Zuverlässigkeit von Direkten Verhaltensbeurteilungen des Lern- und Arbeitsverhaltens in einer inklusiven Grundschulklasse. [Generalizability and Dependability of Direct Behavior Ratings of Academically Engaged Behavior in an Inclusive Classroom Setting]. *Empirische Sonderpädagogik* 7, 258–268.
- Chaffee, R., Briesch, A. M., Johnson, A., and Volpe, R. J. (2017). A Meta-Analysis of Class-Wide Interventions for Supporting Student Behavior. *School Psych. Rev.* 46, 149–164. doi: 10.17105/SPR-2017-0015.V46-2

- Chafouleas, S. M. (2011). Direct Behavior Rating: a Review of the Issues and Research in Its Development. *Educ. Treat. Child.* 34, 575–591. doi: 10.1353/etc.2011.0034
- Chafouleas, S. M., Hagermoser Sanetti, L. M., Kilgus, S. P., and Maggin, D. M. (2012). Evaluating Sensitivity to Behavioral Change Using Direct Behavior Rating Single Item Scales. *Except. Child* 78, 491–505. doi: 10.1177/00144029120780406
- Christ, T. J., Riley-Tillman, T. C., and Chafouleas, S. M. (2009). Foundation for the Development and Use of Direct Behavior Rating (DBR) to Assess and Evaluate Student Behavior. *Assess. Eff. Interv.* 34, 201–213. doi: 10.1177/1534508409340390
- Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., et al. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remed. Spec. Educ.* 36, 220–234. doi: 10.1177/0741932514557271
- Coombes, L., Chan, G., Allen, D., and Foxcroft, D. R. (2016). Mixed-Methods Evaluation of the Good Behavior Game in English Primary Schools. *J. Com. Appl. Social Psych.* 26, 369–387. doi: 10.1002/casp.2268
- DeVries, J. M., Gebhardt, M., and Voß, S. (2017). An Assessment of Measurement Invariance in the 3- and 5-Factor Models of the Strengths and Difficulties Questionnaire: new Insights from a Longitudinal Study. *Pers. Individ. Diff.* 119, 1–6. doi: 10.1016/j.paid.2017.06.026
- Donaldson, J. M., Fisher, A. B., and Kahng, S. W. (2017). Effects of the Good Behavior Game on Individual Student Behavior. *Behav. Anal.* 17, 207–216. doi: 10.1037/bar0000016
- Fabiano, G. A., and Pyle, K. (2019). Best Practices in School Mental Health for Attention-Deficit/Hyperactivity Disorder: a Framework for Intervention. *School Ment. Health* 11, 72–91. doi: 10.1007/s12310-018-9267-2
- Farrell, A. D., Henry, D. B., and Bettencourt, A. (2013). Methodological Challenges Examining Subgroup Differences: Examples from Universal School-Based Youth Violence Prevention Trials. *Prev. Sci.* 14, 121–133. doi: 10.1007/s11212-011-0200-2
- Flower, A., McKenna, J. W., Bunuan, R. L., Muething, C. S., and Vega, R. (2014). Effects of the Good Behavior Game on Challenging Behaviors in School Settings. *Rev. Educ. Res.* 84, 546–571. doi: 10.3102/0034654314536781
- Foley, E. A., Dozier, C. L., and Lessor, A. L. (2019). Comparison of Components of the Good Behavior Game in a Preschool Classroom. *J. Appl. Behav. Anal.* 52, 84–104. doi: 10.1002/jaba.506
- Gebhardt, M., DeVries, J., Jungjohann, J., Casale, G., Gegenfurtner, A., and Kuhn, T. (2019). Measurement Invariance of a Direct Behavior Rating Multi Item Scale across Occasions. *Soc. Sci.* 8:46. doi: 10.3390/socsci8020046
- Goodman, A., Lamping, D. L., and Ploubidis, G. B. (2010). When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): data from British Parents. *Teach. Child. J. Abnorm. Child Psychol.* 38, 1179–1191. doi: 10.1007/s10802-010-9434-x
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a Research Note. *J. Child. Psychol. Psychiatry* 38, 581–586. doi: 10.1111/j.1469-7610.1997.tb01545
- Groves, E. A., and Austin, J. L. (2017). An Evaluation of Interdependent and Independent Group Contingencies During the Good Behavior Game. *J. Appl. Behav. Anal.* 50, 552–566. doi: 10.1002/jaba.393
- Hagermoser Sanetti, L. M., and Kratochwill, T. R. (2009). Treatment Integrity Assessment in the Schools: an Evaluation of the Treatment Integrity Planning Protocol. *Sch. Psychol. Q.* 24, 24–35. doi: 10.1037/a0015431
- Hagermoser Sanetti, L. M., Collier-Meek, M. A., Long, A. C. J., Kim, J., and Kratochwill, T. R. (2014). Using Implementation Planning to Increase Teachers' Adherence and Quality to Behavior Support Plans. *Psychol. Sch.* 51, 879–895. doi: 10.1002/pits.21787
- Hanisch, C., Casale, G., Volpe, R. J., Briesch, A. M., Richard, S., Meyer, H., et al. (2019). Gestufte Förderung in der Grundschule: Konzeption eines mehrstufigen, multimodalen Förderkonzeptes bei expansivem Problemverhalten. [Multitiered system of support in primary schools: Introducing a multistage, multimodal concept for the prevention of externalizing behavior problems]. *Präv. Gesundheitsf.* 14, 237–241. doi: 10.1007/s11553-018-0700-z
- Ho, A. D., and Yu, C. C. (2015). Descriptive Statistics for Modern Test Score Distributions: skewness, Kurtosis, Discreteness, and Ceiling Effects. *Educ. Psychol. Meas.* 75, 365–388. doi: 10.1177/0013164414548576
- Huitema, B. E., and McKean, J. W. (2000). Design Specification Issues in Time-Series Intervention Models. *Educ. Psychol. Meas.* 60, 38–58. doi: 10.1177/00131640021970358
- Hunt, B. M. (2012). *Using the Good Behavior Game to Decrease Disruptive Behavior While Increasing Academic Engagement with a Head Start Population*. Unpublished doctoral dissertation. Hattiesburg: University of Southern Mississippi.
- Joslyn, P. R., Donaldson, J. M., Austin, J. L., and Vollmer, T. R. (2019). The Good Behavior Game: a Brief Review. *J. Appl. Behav. Anal.* 52, 811–815. doi: 10.1002/jaba.572
- Kauffman, J. M., and Landrum, T. J. (2012). *Characteristics of Emotional and Behavioral Disorders of Children and Youth* (10th ed.). Upper Saddle River, NJ: Pearson.
- Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., et al. (2008). Effects of a Universal Classroom Behavior Management Program in First and Second Grades on Young Adult Behavioral, Psychiatric, and Social Outcomes. *Drug Alcohol Depend.* 95, 5–28. doi: 10.1016/j.drugalcdep.2008.01.004
- Klipker, K., Baumgarten, F., Göbel, K., Lampert, T., and Hölling, H. (2018). Psychische Auffälligkeiten bei Kindern und Jugendlichen in Deutschland – Querschnittergebnisse aus KiGGS Welle 2 und Trends. [Mental Health Problems in Children and Adolescents in Germany – Results of the Cross-Sectional KiGGS Wave 2 Study and Trends]. *J. Health Monit.* 3, 37–45. doi: 10.17886/RKI-GBE-2018-077
- KMK (2011). *Inklusive Bildung von Kindern und Jugendlichen mit Behinderungen in Schulen (Beschluss vom 20.10.2011)*. [Resolution of the Standing Conference of the Ministers of Education and Cultural Affairs of the Federal Republic of Germany – Inclusive Education of Children and Adolescents with Disabilities in Schools of October 20, 2011]. Available online at: <https://www.kmk.org/themen/allgemeinbildende-schulen/inklusion.html> (accessed date October 20, 2011)
- Kratochwill, T. R., and Levin, J. R. (2010). Enhancing the Scientific Credibility of Single-Case Intervention Research: Randomization to the Rescue. *Psychol. Methods* 15, 124–144. doi: 10.1037/a0017736
- Lane, K. L., and Menzies, H. M. (2005). Teacher-Identified Students with and without Academic and Behavioral Concerns: Characteristics and Responsiveness. *Behav. Disord.* 31, 65–83. doi: 10.1177/019874290503100103
- Lobo, M. A., Moeyaert, M., Baraldi Cunha, A., and Babik, I. (2017). Single-Case Design, Analysis, and Quality Assessment for Intervention Research. *J. Neurol. Phys. Ther.* 41, 187–197. doi: 10.1097/NPT.0000000000000187
- Maggin, D. M., Pustejovsky, J. E., and Johnson, A. H. (2017). A Meta-Analysis of School-Based Group Contingency Interventions for Students with Challenging Behavior: an Update. *Remedial Spec. Educ.* 38, 353–370. doi: 10.1177/0741932517716900
- Manolov, R., and Moeyaert, M. (2017). Recommendations for Choosing Single-Case Data Analytical Techniques. *Behav. Ther.* 48, 97–114. doi: 10.1016/j.beth.2016.04.008
- Medland, M. B., and Stachnik, T. J. (1972). Good-Behavior Game: a Replication and Systematic Analysis. *J. Appl. Behav. Anal.* 5, 45–51. doi: 10.1901/jaba.1972.5-45
- Moeyart, M., Ferron, J., Beretvas, S., and Van den Noortgate, W. (2014). From a Single-Level Analysis to a Multilevel Analysis of Single-Case Experimental Designs. *J. Sch. Psychol.* 52, 191–211. doi: 10.1016/j.jsp.2013.11.003
- Moore, T. C., Gordon, J. R., Williams, A., and Eshbaugh, J. F. (2022). A Positive Version of the Good Behavior Game in a Self-Contained Classroom for EBD: effects on Individual Student Behavior. *Behav. Disord.* 47, 67–83.
- Mundy, L. K., Canterford, L., Tucker, D., Bayer, J., Romaniuk, H., Sawyer, S., et al. (2017). Academic Performance in Primary School Children with Common Emotional and Behavioral Problems. *J. Sch. Health* 87, 593–601. doi: 10.1111/josh.12531
- Pennington, B., and McComas, J. J. (2017). Effects of the Good Behavior Game Across Classroom Contexts. *J. Appl. Behav. Anal.* 50, 176–180. doi: 10.1002/jaba.357
- Poduska, J. M., and Kurki, A. (2014). Guided by Theory, Informed by Practice: Training and Support for the Good Behavior Game, a Classroom-Based Behavior Management Strategy. *J. Emot. Behav. Disord.* 22, 83–94. doi: 10.1177/1063426614522692
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.
- Saile, H. (2007). Psychometrische Befunde zur Lehrerversion des „Strengths and Difficulties Questionnaire“ (SDQ-L). Eine Validierung anhand soziometrischer Indizes. [Psychometric findings of the teacher version of “Strengths and Difficulties Questionnaire” (SDQ-L). Validation by means of socio-metric indices]. *Z. Entwicklungspsychol. Pädagog. Psychol.* 39, 25–31. doi: 10.1055/s-0033-1343321
- Schulgesetz für das Land Nordrhein-Westfalen (Schulgesetz NRW – SchulG) vom 15. Februar 2005 (GV. NRW. S. 102) zuletzt geändert durch Gesetz vom 23. Februar 2022 (GV. NRW. 2022 S. 250). [School Law for the State of North Rhine-Westphalia

(Schulgesetz NRW - SchulG) of February 15, 2005 (GV. NRW. p. 102) last amended by the Act of February 23, 2022 (GV. NRW. 2022 p. 250)].

Shadish, W. R., and Sullivan, K. J. (2011). Characteristics of Single-Case Designs Used to Assess Intervention Effects in 2008. *Behav. Res.* 43, 971–980. doi: 10.3758/s13428-011-0111-y

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Rindskopf, D. M., Boyajian, J. G., and Sullivan, K. J. (2014). “Analyzing Single-Case Designs: d, G, Hierarchical Models, Bayesian Estimators, Generalized Additive Models, and the Hopes and Fears of Researchers about Analyses,” in *Single-Case Intervention Research. Methodological and Statistical Advances*, eds T. R. Kratochwill and R. L. Levin (Washington DC: American Psychological Association), 247–281.

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., and Sullivan, K. J. (2013). Bayesian Estimates of Autocorrelations in Single-Case Designs. *Behav. Res. Methods* 45, 813–821. doi: 10.3758/s13428-012-0282-1

Simonsen, B., and Myers, D. (2015). *Classwide Positive Behavior Interventions and Supports. A Guide to Proactive Classroom Management*. New York, NY: Guilford Press.

Smith, R. L., Eklund, K., and Kilgus, S. P. (2018). Concurrent Validity and Sensitivity to Change of Direct Behavior Rating Single-Item Scales (DBR-SIS) within an Elementary Sample. *Sch. Psychol. Q.* 33, 83–93. doi: 10.1037/spq0000209

Smith, S., Barajas, K., Ellis, B., Moore, C., McCauley, S., and Reichow, B. (2021). A Meta-Analytic Review of Randomized Controlled Trials of the Good Behavior Game. *Behav. Modif.* 45, 641–666. doi: 10.1177/0145445519878670

Spilt, J. L., Koot, J. M., and van Lier, P. A. C. (2013). For Whom Does it Work? Subgroup Differences in the Effects of a School-Based Universal Prevention Program. *Prev. Sci.* 14, 479–488. doi: 10.1007/s11211-012-0329-7

Tanol, G., Johnson, L., McComas, J., and Cote, E. (2010). Responding to Rule Violations or Rule Following: a Comparison of Two Versions of the Good

Behavior Game with Kindergarten Students. *J. Sch. Psychol.* 48, 337–355. doi: 10.1016/j.jsp.2010.06.001

Van den Noortgate, W., and Onghena, P. (2003). Combining single case experimental data using hierarchical linear modeling. *Sch. Psychol. Q.* 18, 325–346. doi: 10.1521/scpq.18.3.325.22577

van Lier, P. A. C., Vuijk, P., and Crijnen, A. A. M. (2005). Understanding Mechanisms of Change in the Development of Antisocial Behavior: The Impact of a Universal Intervention. *J. Abnorm. Child. Psychol.* 33, 521–535. doi: 10.1007/s10802-005-6735-7

Volpe, R. J., Casale, G., Mohiyeddini, C., Grosche, M., Hennemann, T., Briesch, A. M., et al. (2018). A Universal Screener Linked to Personalized Classroom Interventions: Psychometric Characteristics in a Large Sample of German Schoolchildren. *J. Sch. Psychol.* 66, 25–40. doi: 10.1016/j.jsp.2017.11.003

Voß, S., Blumenthal, Y., Mahlau, K., Marten, K., Diehl, K., Sikora, S., et al. (2016). *Der Response-to-Intervention-Ansatz in der Praxis. Evaluationsergebnisse zum Rügener Inklusionsmodell*. [The Response-to-Intervention Approach in Practice. Evaluation Results on the Rügen Inclusion Model]. Münster: Waxmann.

Wang, M.-T., and Eccles, J. S. (2013). School Context, Achievement Motivation, and Academic Engagement: a Longitudinal Study of School Engagement Using a Multidimensional Perspective. *Learn. Instr.* 28, 12–23. doi: 10.1016/j.learninstruc.2013.04.002

Waschbusch, D. A., Breaux, R. P., and Babinski, D. E. (2019). School-Based Interventions for Aggression and Defiance in Youth: a Framework for Evidence-Based Practice. *School Ment. Health* 11, 92–105. doi: 10.1007/s12310-018-9269-0

Wilbert, J., and Lüke, T. (2018). *Scan: Single-Case Data Analyses for Single and Multiple Baseline Designs*. Available at: <https://r-forge.r-project.org/projects/scan> (accessed date 2016-06-23).

Wiskow, K. M., Ruiz-Olivares, R., Matter, A. L., and Donaldson, J. M. (2018). Evaluation of the Good Behavior Game with a Child with Fetal Alcohol Syndrome in a Small-Group Context. *Behav. Interv.* 33, 150–159. doi: 10.1002/bin.1515



OPEN ACCESS

EDITED BY

Sarah Powell,
University of Texas at Austin,
United States

REVIEWED BY

Lénia Carvalhais,
Infante D. Henrique Portucalense
University, Portugal
Ingrida Balėiūnienė,
Vytautas Magnus University, Lithuania

*CORRESPONDENCE

Sandra L. Gillam
sandi.gillam@usu.edu

SPECIALTY SECTION

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

RECEIVED 12 April 2022

ACCEPTED 21 July 2022

PUBLISHED 08 August 2022

CITATION

Israelsen-Augenstein M, Fox C,
Gillam SL, Holbrook S and Gillam R
(2022) Monitoring indicators
of scholarly language: A progress
monitoring tool for documenting
changes in narrative complexity over
time.
Front. Educ. 7:918127.
doi: 10.3389/feduc.2022.918127

COPYRIGHT

© 2022 Israelsen-Augenstein, Fox,
Gillam, Holbrook and Gillam. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Monitoring indicators of scholarly language: A progress monitoring tool for documenting changes in narrative complexity over time

Megan Israelsen-Augenstein¹, Carly Fox², Sandra L. Gillam^{2*},
Sarai Holbrook³ and Ronald Gillam²

¹Communication Sciences and Disorders Division, West Virginia University, Morgantown, WV, United States, ²Communicative Disorders and Deaf Education, Utah State University, Logan, UT, United States, ³Communication Sciences and Disorders, University of Wisconsin – Stevens Point, Stevens Point, WI, United States

The purpose of this cross-sectional study was to determine the differences in narrative macrostructure abilities of children in different age groups using a progress monitoring tool based in discourse theory. A majority of existing research regarding narrative developmental patterns has been based in schema theory. The *Monitoring Indicators of Scholarly Language* (MISL) rubric is based in discourse theory and was designed to characterize aspects of narrative proficiency in school-age children. The data for this project consisted of 687 narratives elicited using the Aliens subtest from The *Test of Narrative Language—Second Edition* (TNL-2). There were 1,597 participants who ranged in age from 4; 0 to 15; 0 (year; month). An ordinary least squares regression where age predicted total macrostructure score, followed by a series of *post hoc* ordinal logistic regressions (OLR) where age predicted each individual MISL rubric element was used. Results of both the simple regression on total macrostructure score and the series of ordinal regression analyses for each macrostructure element indicated that age was a significant predictor of the scores children received. Collectively, these results suggest that the MISL is a developmentally valid measure of narrative production abilities. Developmental milestones based on discourse theory are reported to be substantially later than has been reported for schema theory. The differences are highlighted and the implications for progress monitoring for narrative development are discussed.

KEYWORDS

narrative language, progress-monitoring measures, narrative macrostructure, language impairment, narrative discourse

Introduction

The study of narrative discourse is a critical pursuit in the field of speech language pathology, particularly for professionals who work with school-aged children. Discourse can be defined as text or spoken language beyond the sentence level (Hughes et al., 1997; Nicolosi et al., 2004), while narratives are a genre of discourse also known as stories (Berman and Nir-Sagiv, 2007; Graham et al., 2013; Dockrell et al., 2014). Knowledge and use of narrative discourse requires a child to produce stories that contain specific structural features of narrative language and serves a specific communicative goal (Berman and Nir-Sagiv, 2007; Carvalhais et al., 2021). Narrative discourse is valued in the study of school-aged children's language because the ability to successfully produce a narrative is considered an important developmental milestone and is included in the Common Core State Standards for students in the United States (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010). In addition, research has suggested that preschool and school-age children who struggle with narrative production and comprehension are more likely to experience later academic difficulties in tasks involving reading, writing, and oral language (Liles et al., 1995; Greenhalgh and Strong, 2001; Catts et al., 2002; Roth et al., 2002; Justice et al., 2006; Gillam et al., 2017). Narrative comprehension and production require a complex integration of social, linguistic, pragmatic, and cognitive skills that make it an ideal method for studying a child's communication abilities (Liles, 1993; Wagner et al., 2000; Botting, 2002; Nippold et al., 2014). Due to their complex nature, narratives can and are used as a measure of language ability for students in a wide age range (MacLachlan and Chapman, 1988; Dollaghan et al., 1990; Leadholm and Miller, 1992; Wagner et al., 2000; Westerveld et al., 2004; Nippold et al., 2014). While most researchers agree that typically developing children produce "adult-like" narratives by the age of six or seven (Hughes et al., 1997), there is evidence that narratives continue to grow in complexity through adolescence (Applebee, 1978; Peterson and McCabe, 1983; Roth and Speckman, 1986; Purcell and Liles, 1992; Liles, 1993; Crais and Lorch, 1994; Munoz et al., 2003; Stadler and Ward, 2005).

This makes the evaluation of narrative discourse skill, which is often conducted through language sample analysis, a unique context in which to gain a more complete picture of a child's language profile over time. Therefore, a number of progress monitoring tools have been designed to make discourse level analysis more accessible to speech language pathologists. These tools have largely been based in schema theory, which has been the prevailing conceptualization of narrative discourse structure used in the field of speech language pathology and education. In schema theory, narratives are largely defined by the elements they contain (Meyer, 1975; Mandler and Johnson, 1977; Rumelhart, 1977; Thorndyke, 1977; Stein and Glenn, 1979) and consist of a setting and episode (Stein and Glenn, 1979). The

setting system consists of the main character(s) and the physical, temporal and/or spatial location of the story. An episode includes the main character, a goal, actions/attempts directed at achieving that goal (often referred to as attempts), and a consequence or resolution. More complex episodes include the main character's internal responses (feelings) related to the goal, and plan(s) to achieve their goal. Aspects of the story (character, setting, goals, plans, actions, consequences) are referred to as story grammar elements (SGEs).

Schema theory views a child's ability to comprehend or produce a narrative as related to their internal organization or knowledge of story grammar elements (Rumelhart, 1975; Mandler and Johnson, 1977; Stein and Glenn, 1979). Schemata, in a sense, form our expectations of what components a story should possess, so that information can be processed more efficiently. Stein and Glenn (1979) proposed developmental stages for the types of narratives that children produce with preschoolers often telling stories that "describe characters" and "list actions" in a temporal order. At about 6 years of age, children are said to tell stories that include the aims or intentions of the character but may not include specific names of their characters. Between the ages of 7–8, children were reported to begin telling stories that have a "chain of reactive sequences" or "abbreviated episodes." Key elements included in an abbreviated episode include an initiating event with a chain of actions taken by the characters. Stories that include an abbreviated episode often omit a conclusion.

A full episode, according to Stein and Glenn (1979), contains a basic episode and is produced by children 8–9 years of age. A story that includes a full episode includes an initiating event, attempts, and a consequence that are related and included in a cohesive and sequential order. By the time a child is 11, they are said to tell stories that are complex and include elaborated episodes with multiply embedded plans, and/or attempts.

A longitudinal study conducted by Berman and Nir-Sagiv (2007) is one of the only studies that has been conducted to substantiate the developmental stages proposed by Stein and Glenn (1979). The researchers analyzed written texts produced by 80 English-speaking children and adults. The participants were split into four age groups (i.e., elementary school children, junior high school students, high school students, and university students). Each participant was asked to write a narrative retell from a video prompt. These stories were then coded for different linguistic and narrative elements to determine the complexity of the narratives produced by individuals in each age group. They reported that story retells produced by the elementary school children in this study differed from reports of stories produced by preschool children. Preschool children typically produce stories that often contain weakly developed narrative macro- and microstructure elements. For example, the children do not produce a wide variety of macrostructure elements (i.e., story grammar elements; SGEs), and fail to include microstructure elements like subordinating and coordinating conjunctions. The

researchers proposed that the preschool children in their studies may not have had the linguistic and cognitive skills necessary to establish causal and temporal relationships within their narratives. However, by early school-age (around 4th grade, or approximately ages 8–9), they determined that the children were developing linguistic foundations that allowed them to include basic story elements in their written stories. These children did not yet produce narratives that were elaborate and included clear causal connections. The data gathered from this study indicated that children may develop the ability to produce well-formed narratives that utilize all SGEs after fourth grade. The researchers' findings differed from [Stein and Glenn \(1979\)](#) because it wasn't until children were older (i.e., junior high school students) that they produced written narratives that were well developed and included complex and elaborated episodes. These findings suggest that oral and written narrative development may differ slightly, however, similar development patterns are observed. Knowing the trajectory of narrative development can assist speech-language pathologists in applying and understanding the needs of children when using progress-monitoring tools to assess narrative abilities of children on their caseloads.

Assessment tools based in schema theory require the examiner to note the presence or absence of specific story elements used by the storyteller and have been elicited from a range of story prompts (e.g., story retells, sequenced pictures). Children who include certain elements, or "more" story elements are thought to have better narrative abilities than children who omit important story elements or use fewer elements ([Berman, 1988](#); [Strong, 1998](#); [Boudreau and Hedberg, 1999](#); [Miles and Chapman, 2002](#); [Reilly et al., 2004](#)). Other scoring rubrics incorporate subjective "text-level" judgments to rate overall story quality ([Applebee, 1978](#); [Stein, 1988](#); [Hedberg and Westby, 1993](#)).

For example, The Strong Narrative Assessment Procedure (SNAP; [Strong, 1998](#)) is a tool aligned with schema theory designed to measure both macro- and microstructure elements of narratives. Standardized samples of text-level discourse are elicited using audiotaped stories that narrate the wordless picture books: A Boy, a Dog and a Frog ([Mayer, 1967](#)), Frog, Where Are You? ([Mayer, 1969](#)), Frog Goes to Dinner ([Mayer, 1974](#)), and One Frog Too Many ([Mayer, 1975](#)). Children are asked to retell each of these stories after listening to them. The story retells are recorded, transcribed and analyzed for 26 different narrative macrostructure and microstructure elements. This assessment provides information about overall use of SGEs and general language features in the stories children produce. Both of these assessments provide information related to the knowledge and use of specific types of SGEs used by children in their stories. Differences in narrative performance across ages have been documented in narrative retell tasks using this assessment and are comparable to the developmental data reported by [Stein and Glenn \(1979\)](#) and [Berman and Nir-Sagiv \(2007\)](#) (see [Strong, 1998](#); [John et al., 2003](#)).

[John et al. \(2003\)](#) completed a study to determine if the SNAP yielded differences in story retelling abilities in children across different ages. Story retell samples were elicited from 61 typically developing children between the ages of 6 and 11. The children were assigned to three different age groups for the purpose of data analysis. The SNAP assessment (i.e., story grammar analysis) was used to score the story retell samples that were elicited using four wordless picture books created for the study. The researchers found that the mean scores for the proportion of story grammar elements retold were consistent with previous literature (e.g., [Stein and Glenn, 1979](#)) because the children in their sample recalled initiating events, attempts, and consequences more often than elements like internal response. Age was found to be a significant predictor for the story grammar element of internal response. The children in the youngest age group recalled significantly fewer instances of internal response as compared to the other two age groups. In addition, the children in the oldest age group (i.e., 11-year-olds) reported the element of internal response more often than those in either of the other groups. Children in this sample were demonstrating the use and recall of elements included in a basic story episode (i.e., initiating event, attempt, and consequence), as well as recalling instances of internal response by the time they were 7 years old. As children grew older, they demonstrated a greater number of recalls of instances of internal response, with children who were 11 years old including a significantly larger amount than all other children included in the study. Using the SNAP, a story grammar analysis is gathered that has been shown to yield differences in narrative story retelling ability across age. However, information regarding the causal and temporal connections established by children in stories and the global organization and coherence of a story is not gleaned using this assessment. This information may be necessary in order to more completely understand a child's narrative discourse abilities.

An alternate understanding of narrative ability comes from discourse theory, which is a broad subfield of linguistics dedicated to the study of language and communication beyond the sentence level. Here, we use the term discourse theory to refer to the narrower scope of discourse processing, as described in the Event-Indexing model, which focuses on the construction of a mental representation during narrative comprehension and production ([Zwaan et al., 1995](#)). Discourse theory can be thought of as an extension of schema theory that aims to account for a larger number of variables. Discourse theorists recognize the importance of schemata in narrative production and comprehension, as schemata as it allows for the child to more efficiently and accurately recall story information and adapt/monitor their mental representation of the story content ([Kintsch and van Dijk, 1978](#); [van Dijk and Kintsch, 1983](#); [Zwaan et al., 1995](#)). However, discourse theory provides a more in-depth explanation for how local and global coherence in narratives (or more generally discourse) are established. [Kintsch and van Dijk \(1978\)](#) highlighted the importance of

forming a coherent text-base in order to create a fully formed situation (mental) model of the narrative. They proposed that when processing a narrative, the individual clauses are reviewed at the proposition level (i.e., a predicate and argument) and are compared for argument overlap within working memory. From this point of view, the likelihood that a proposition is stored and subsequently influences the mental representation of the narrative increases as a function of the amount of times they overlap with other propositions across causal (i.e., the causal connections between events) and temporal (i.e., the temporal relationship between events) dimensions (Zwaan et al., 1995). The greater the overlap, the more likely a particular proposition is to be important to the central theme or plot. The explicit inclusion of causal and temporal connections between events in a story is a critical component of evaluation for discourse theory-based narrative assessment tools, as it provides an objective measure of the storyteller's understanding of the relationship between events in a story. Therefore, narrative progress monitoring tools based in discourse theory require a measurement of the level and overlap of propositions across causal and temporal events in a story in addition to a measurement of the use and knowledge of SGEs. Notably such measures of causal and temporal events in a story are not seen in tools designed from schema -theory.

One of the first tools to incorporate both aspects (discreet scores for story elements and holistic ratings of overall story quality) was the Narrative Scoring Scheme (NSS; Miller et al., 2003; Heilmann et al., 2010). The Narrative Scoring Scheme was developed to measure the use of specific story elements as well as overall story "quality" using story retelling of Frog Stories (e.g., Mayer, 1967, 1969, 1974, 1975). Key story elements measured include an introduction, conflicts, and the conclusion of the story. These elements constitute the "macrostructure" analysis, whereby other aspects of language microstructure are measured by noting the presence or absence of language used to describe character development and to differentiate between the main and supporting characters. Holistic judgments are also made to analyze inter-textual cohesive quality referencing, and cohesion.

To our knowledge, no study has evaluated the development of children's oral narrative abilities using a rubric designed specifically to measure aspects of macrostructure and microstructure based in discourse theory. As discourse theory transcends the use of schema theory, it may be beneficial to understand the developmental trajectory of children's narratives in their school-age years. The *Monitoring Indicators of Scholarly Language* (MISL; Gillam et al., 2017) rubric is a progress-monitoring tool that has been designed to track children's development of oral narrative skill over time and is based in discourse theory. The MISL was designed to measure stories that range from simple descriptions to complex multi-episodic narratives. Both a macrostructure and microstructure subscale are included and yield a total narrative proficiency score based in discourse theory. The macrostructure subscale accounts for the

use of SGEs as well as the level to which each element is causally and temporally connected in the global organization of a story. The microstructure subscale accounts for the use of literate language features necessary to establish temporal and causal connections locally in stories. It utilizes discrete measurement criteria for the use of story grammar elements as well as the causal connections between them reflecting the nature of the interrelationships between critical episodic elements. This is achieved by removing some of the subjectivity inherent in the use of holistic judgments and making them "discreet." This is more in line with the notion of macrostructure as introduced by Kintsch and van Dijk (1978) (and others), who maintain that macrostructure is not measured by documenting the presence or absence of story elements or holistic judgments of cohesion, but rather by the causal framework that exists between them.

The MISL has been shown to be a valid and reliable measure for charting progress in oral narrative growth (Gillam et al., 2017). In the first study, the MISL was used to score stories told by 109 children with language impairments (ages 5; 7–9; 9) who participated in a normative study for the Test of Narrative Language—Second Edition (TNL-2; Gillam and Pearson, 2017). The stories elicited from the Aliens subtest were used to assess psychometric adequacy measured for inter-rater reliability, internal consistency and construct validity. The Aliens subtest is an oral narrative prompt where the child is asked to tell a story from a picture. The MISL was shown to demonstrate good inter-rater reliability for the macrostructure and the microstructure subscales (ranging from 92 to 100% for each item) and acceptable levels of both internal consistency reliability (>0.70 Cronbach's alpha) and construct validity for use in measuring overall narrative proficiency (MISL total score). It has yet to be established, however, whether each subscale element is developmentally sensitive to narrator age and if so, whether that extends beyond the elementary school-age range (5–9 years). In addition, we were interested in knowing whether MISL scores across ages reflect the same developmental stages proposed by Stein and Glenn (1979) that were supported by Berman and Nir-Sagiv (2007).

Measurement of the presence of SGEs as well as their causal relationship to one another is critical if we are to gain a more thorough understanding of a child's knowledge of narrative structure. Research has explored the role of causal connectivity in written discourse and has revealed that statements in written text that include a large number of causal connections tend to be more readily recalled (Espin et al., 2007), judged as more important by the reader (Trabasso and Sperry, 1985), and retrieved from memory more quickly (O'Brien and Meyers, 1987) than statements that have a smaller number of connections. Similar findings have been described in oral discourse tasks (Cevasco and van den Broek, 2008). Though previous summaries of child's narrative development (Stein and Glenn, 1979; Berman and Nir-Sagiv, 2007) have

reported children incorporating basic episodes in their stories around ages of 7–8 years old, we predict that MISL scores would reflect a later timeline, as they require the narrator to explicitly indicate the causal relationships between SGEs, not just simply state events in a logically ordered sequence (as is the case in schema theory-based assessment tools). We therefore hypothesize that children in this age range who may have less knowledge of narrative structure, as well as less well-developed language abilities, may frequently fail to produce basic episodes with explicitly stated causal connections between initiating events, actions and consequences. To determine the typical age at which both basic and elaborate narratives were produced based on discourse theory criteria, narratives produced by children from a larger age range, including older school-age children (4–15 years), were evaluated using the MISL rubric.

The purpose of this project was to understand the nature of child's oral narrative development following discourse theory using a progress monitoring tool for children ages 4–15. To address this purpose, the following questions were posed:

1. Are measurements of macrostructure ability (as measured by the MISL rubric) sensitive to changes across age?
2. At what age do the majority of children in the sample achieve proficiency (i.e., a score of 2 or more) across each macrostructure element?

Materials and methods

Participants

A total of 687 narratives were analyzed in this study, which were elicited from participants drawn from the normative sample of 1,597 children in the TNL-2 (Gillam and Pearson, 2017). The participants ranged in age from 4; 0 to 15; 0 ($M = 8; 9$, $SD = 2; 8$). There was a roughly even split in the reported biological sex of participants, with 51.8% of narratives ($n = 356$) having been elicited from females, and 48.2% of narratives ($n = 331$) having been elicited from males. Samples were elicited from children whose reported ethnicity was white (86%), black or African American (9%), two or more ethnicities (2%), Asian or Pacific Islander (1.5%), American Indian, or Alaskan Native (0.6%), with the remaining 0.9% preferring to not respond. Close to one-third of the sample were identified as qualifying for free and reduced lunch programs (29.1%), with the remaining 70.9% either not qualifying or choosing to not report this information. Finally, narratives were elicited from children across different regions in the United States, including the Northeastern region (21.8%), the Southeastern region (16.7%), the Midwestern region (7.6%), and the Western region (53.9%).

Materials

The TNL-2 (Gillam and Pearson, 2017) is a standardized measure of narrative proficiency that assesses a child's comprehension and production of stories in three progressively independent contexts. The first context requires participants to listen to a story, answer questions about it, and retell the story (McDonald's subtest). Next, students are asked to listen to a story that is modeled using a set of sequenced pictures (Shipwreck subtest), answer questions about it, and then create a new account with a novel set of sequenced images (Late for School subtest story). The last context involves asking participants to listen to a story about a single picture (Treasure subtest), answer questions about it, and create a new account from a unique image (Aliens subtest). The prompt for the Aliens subtest is a novel scene that depicts an alien family that is landing in the park. Children are asked to generate a story based off of the picture prompt. The narratives for this project were elicited from the Aliens subtest of the TNL-2 assessment.

The MISL rubric was used to score the Aliens subtest story from each participant's TNL-2 assessment. The MISL includes a macrostructure and microstructure subscale. The scores from these scales are then combined to reflect an overall narrative proficiency score. Story elements are judged as absent (score of 0), emerging (score of 1), present/mastered (score of 2), or elaborated (score of 3). Scores on the MISL are awarded based on how the story elements (e.g., initiating event, action, consequence) are causally/temporally related rather than the number of times an element is observed in a narrative. A score of 0 is interpreted as evidence that the story does not contain the elements that make up a basic story episode. These stories may contain simple descriptions of objects or actions (e.g., There is a ship. They are eating). A score of 1 indicates that a story may have an emerging episodic structure (e.g., There is a girl. She is hiding in the bush). A score of 2 is interpreted as evidence that a story contains the necessary elements to constitute a basic story episode (e.g., The girl is hiding behind a bush and then jumped out to scare the aliens. She ran home to tell her parents about the aliens because she was scared). A score of 3 indicates that the story is complex and elaborated (e.g., Jill and Jack were at the park. They hid behind the bush because the aliens landed. They decided to jump out from behind the bush to scare the aliens. After they scared the aliens, they ran home to tell their parents all about their day at the park. Their parents didn't believe their story, so they took them back to the park. When they got to the park, the aliens were gone). The macrostructure subsection of the MISL is designed to measure both SGEs and the temporal and causal connections that make the narrative both locally and globally coherent. There are seven SGEs measured in the MISL, including Character, Setting, Initiating Event, Internal Response, Plan, Action, and Consequence (see Table 1). Similar to the view of Stein and Glenn (1979), Character and Setting are scored individually, as they exist outside of the overall sequence

of the plot. The remaining elements comprise a chain of events that begins with Initiating Event and resolves with consequence. This causal chain is critical to maintaining the global cohesion of a narrative that allows the story recipient to construct and maintain a situation model, or a mental representation of the narrative, which underlies narrative comprehension (Zwaan et al., 1995; Graesser et al., 1997).

In order to determine whether a causal connection exists between statements in a story, a cause must come before its outcome (temporal priority), be in operation when the outcome occurs (operativity) and be necessary for the consequence to occur (necessity; Mackie, 1980; van den Broek, 1990; Zwaan et al., 1995). Children often produce stories in which the conditions for causality are not met. For example, in the story, “John went to the store to buy some food. He forgot his money.” It is implied that John was unable to buy food because he did not bring his money. The conditions necessary for causality are not met in this case because, while there is a temporal order (went to store, forgot money) there is no “outcome” stated. Most narrative macrostructure scoring systems, that are based in schema theory, would not only the presence or absence of specific story elements with “more” being better than less (Berman, 1988; Strong, 1998; Boudreau and Hedberg, 1999; Miles and Chapman, 2002; Reilly et al., 2004).

While the presence or absence of story elements is part of the MISL scoring system, it also includes judgments about the causal nature of the events in the story. The conditions of causality in scoring story episodes is reflected in MISL scoring by utilizing an interdependent scoring system between initiating event, internal response, plan, action, and consequence. The minimal score that indicates the conditions of causality are met is a 2 for each of these items. For example, if the story stated:

John went to the store to buy groceries. He forgot his money, so he was not going to be able to buy his food. He decided to call his mother and ask her to bring him some money so he would be able to buy his groceries. He called his mother, and she was happy to bring him some money. After John’s mom brought him money, he finished his grocery shopping and came home to make his mom dinner to thank her for saving the day.

The initiating event in the story was the problem of John not being able to buy food without money. He then called his mother [action causally related to buying groceries (initiating event)], requested funds (action), and received funds (action). John then bought groceries (consequence) because that is what he originally came to the store to do (initiating event). Schema systems might give credit for the presence or absence of the initiating events, action, and consequences because they are stated in the story. For example, an action might be identified if a story contained the sentence “The girl ran over to her mother.” However, in order for this statement to earn a score of 2 for attempt using the MISL, it would need to be clearly tied to an initiating event such as, “The girl ran over to her mother because she was afraid of the thunder.” In the previous sentence, a score

of 1 would be given for the sentence, “The girl ran over to her mother” using the MISL because there is no “clear link to an initiating event” using causal language.

General procedures

Story transcription

Stories were recorded on portable digital audio recorders and transcribed verbatim by research assistants who were blind to the purpose of the study. *Systematic Analysis of Language Transcripts* conventions were used to code each utterance (SALT; Miller and Iglesias, 2019). The utterances were segmented into communication units (C-units) consisting of an independent main clause and phrases or clauses subordinated to it. Each transcript was reviewed by a second research assistant for spelling, mazing, morpheme segmentation and utterance segmentation. Transcription disagreements were addressed by both transcribers who listened to the digital recording together and discussed the differences until a resolution was reached. Reliability between primary and secondary transcribers was calculated on 20% of the data. The total number of C-units and mazes (i.e., false starts, revisions) were calculated, and the number of discrepancies were determined. The discrepancies were then subtracted from the total number of C-units and mazes and a percentage agreement was calculated. Reliability was 96.7% for C-unit segmentation and 96.1% for identification and coding of mazes.

Monitoring indicators of scholarly language training

Research assistants met with the first author to review the subscales, definitions and scoring criteria of the MISL using example stories. Twenty stories that represented a variety of story types and quality levels were selected for use in MISL training. Research assistants were given five stories at a time to score. After they were scored, the research assistants met with the first author to discuss the scores and the reasoning behind the scoring decisions. This process was repeated until all 20 had been scored. After the training period, the research assistants were given 10 new stories to score that were not part of the TNL-2 database. These stories were used to determine when a research assistant had reached an overall and point-by-point reliability score of 80% or higher for scoring the MISL subscales. Only then were they considered to be sufficiently trained to participate in scoring stories for the study.

Two research assistants who had met these criteria and who were blind to the purpose of the study independently used the MISL rubric to score de-identified narrative transcripts. These research assistants independently scored stories in increments of 30. After each subset of 30 stories, the research assistants met together with the first author to review scores and discuss any scoring disagreements. This was done to minimize any effect of

TABLE 1 Macrostructure subscale story elements and scoring criteria.

Story element	0 (not present)	1 (emerging)	2 (mastery)	3 (elaborated)
Character	No main character is included or an ambiguous pronoun is used to reference a person	Includes at least one main character by using a non-specific label with a determiner (e.g., the boy, a girl)	Includes at least one main character that is referenced to using a proper noun	Includes more than one main character using proper nouns
Setting	No reference to a location or time is used	Only references to a general place or time is included (this reference is not necessarily related to the story)	Reference to a specific time or place that is related to the story is included	A reference to the place are created using proper nouns, and a reference to a specific time are included
Initiating Event	No indication of an initiating event—series of descriptions	Initiating event is stated, however, this event does not motivate actions from the characters	One initiating event is stated that motivates actions from the main characters	Two or more initiating events are included that motivate separate actions from the main characters
Internal Response	No feelings from the characters are stated	Feelings from the characters are stated, however, there is not clear relationship to the initiating event.	Feelings are stated that is clearly related to the initiating event	Multiple instances of feelings are stated that are clearly related to the initiating event.
Plan	No statement is included that describes the character's plan to take action	Statements about plans to take action are included, however, these plans are not directly related to the initiating event.	One statement depicting a plan is included that is directly related to the initiating event.	Multiple statements about plans the characters have to take action are included that are directly related to the initiating event.
Attempt	No actions/attempts are taken by the characters	There is use of action verbs in descriptive sequences that do not have a clear link to an initiating event.	The use of action verbs in the story are clearly linked to the initiating event	A complicating action that impedes the actions characters take in response to the initiating event are included.
Consequence	There is no clear “ending” or resolution stated that is related to an initiating event	The outcome or resolution of the action is linked to another action, not the initiating event	One resolution of actions stated that is directly related to the initiating event	Two or more outcomes are stated that are directly related to the initiating event

coder drift, which is a phenomenon resulting from systematic and predictable variation in rater decisions over time. Any differences in scores were discussed and resolved by the research assistants under the direction of the first author. Reliability on each macro- and microstructure element was calculated on the uncorrected data for each item (point by point). The number of agreements was divided by the total number of item decisions and then multiplied by 100. Reliability between primary and secondary scorers was calculated on 100% of the data for the project. Interrater reliability for MISL total scores was 85%.

Data analysis

Pearson correlation analysis was used to first establish convergent validity between the macrostructure section of the MISL rubric and the TNL-2 Aliens subtest raw production score. This step was necessary to establish the appropriateness of utilizing the normative database collected for the TNL-2 as a normative database for MISL scores. The macrostructure total score and the TNL-2 Aliens subtest raw production scores were found to have high levels of convergent validity, based on correlation analysis, $r(686) = 0.766$, $p < 0.001$, indicating that the normative sample for the TNL-2 could adequately serve as a normative sample for the MISL macrostructure.

Research question one, which aimed to determine the sensitivity of each macrostructure element to the age of narrator was addressed through an ordinary least squares regression where age predicted total macrostructure score, followed by a series of *post hoc* ordinal logistic regressions (OLR) where age predicted each individual MISL rubric element. OLR was utilized to capture the ordinal nature of the MISL scores, which are on a scale of 0–3, with each score representative of a different level of narrative proficiency. Use of a generalized linear modeling method like OLR was necessary, as both ordinary least squares and analysis of variance assume a continuous dependent variable with normally distributed residuals. In each OLR, age of narrator predicted each individual macrostructure element score (Character, Setting, Initiating Event, Plan, Internal Response, Action, Consequence) for a total of seven models. Beta coefficients were converted to odds-ratios for ease of interpretation.

To address research question two, which was to evaluate the age at which the majority of the children in the sample had proficient scores (i.e., a score of two or higher) for each macrostructure element, descriptive statistics were utilized. Mainly an evaluation of the modal score for each age (separated by year) was evaluated for each macrostructure element.

Results

The ordinary least squares linear regression indicated that age of narrator was a significant predictor of total macrostructure score $\beta = 0.97$, $t(687) = 19.52$, $p < 0.001$, meaning that each 1-year increase in age was associated with a 0.97 point increase in macrostructure total score (see Table 2). The R-squared value estimates that 35.64% of the variance in macrostructure total score can be accounted for by age of the narrator.

Results of each OLR model indicated that age was a significant predictor of all macrostructure elements ($p < 0.001$), whereby a positive trend was seen between the age of narrator and their score on each of the seven macrostructure elements. Odds-ratios ranged between 1.13 and 1.60, indicating that for each 1-year increase in age of narrator, the odds of receiving the next highest macrostructure score increased by 1.13–1.60 times across each of the elements. The smallest effect size was seen for internal response, however, the relationship between age of narrator and MISL score was still statistically significant, [ordered odds ratio (Estimate)] = [1.13], 95% CI = [1.07, 1.19], Wald = [4.522], $p < 0.001$. The largest effect size was seen for Consequence, where each 1-year increase in age was associated with 1.6 times increase in the odds of receiving the next score level, 95% CI = [1.50, 1.71], Wald = [13.975], $p < 0.001$. Results of each OLR model are presented in Table 3.

An array of Jitter plots depicting the distribution of individual scores for each element by age is shown in Figure 1. Modal scores (i.e., the most commonly occurring score) by age for each element are discussed in the following sections and are also depicted in Table 4. Modal scores are provided in place of the mean and standard deviation, since scores are

ordinal in nature and represent different stages of SGE mastery. Scores were also not normally distributed, so the mean score for each age would not accurately represent the middle of the score distribution.

Examination of the distribution of scores in Figure 1, revealed a distinct increase in scores for character for 9- and 10-year-old children; whereby younger children ages 4–9 most frequently received (mode) a score of 1 on character. Children between the ages of 10–15 most frequently received (mode) a score of 3 for character, indicating not only proficiency for this age range, but elaboration.

The modal value for setting remained at a score of 1 across all ages, however, it can be seen in Figure 1 that the distribution of scores was more widespread from ages 10 on. This means that while 1 remained the most common setting score regardless of age, older children were more likely to include Setting at the proficient or elaborated level.

For initiating event score the modal score was consistently 2, indicating proficiency, for ages seven and older. The Jitter plot in Figure 1 shows an evident cluster of scores at 3 for initiating event from age 8 and older, and a cluster of scores 0 and 1 for ages 4–7, with 2 remaining the most frequent initiating event score across all ages.

The modal score for internal response was 0 for each age apart from the 12 and 15-year-old group. As can be seen in the Jitter plot for internal response in Figure 1, the largest cluster of scores across ages was 0, however, there was a smaller number of scores at 0 from ages 12 on. This finding indicated that while it was common for narrators to exclude the use of internal response in their stories, there was greater likelihood for its inclusion at later ages.

For plan, there was a clear increase at ages 9 and 10 in its presence and sophistication in children's stories. Prior to that, for ages 4–9 the most frequent score for plan was 0. By the time students reached ages 10–15 the most frequent score for plan was 2, indicating proficiency in using the story element causally to indicate intentions of characters. The Jitter plot of plan in Figure 1 reflects these clusters, in addition to showing a small cluster of scores at 1 for the middle age range and a sparse cluster of scores at 3 in the older age range.

Following plan, action had a modal score of 2 across the majority of the age-range included in the sample (7; 0–15; 0). As can be seen in the Jitter plot, there appeared to be a greater spread in scores for narratives elicited from children between 5; 0 and 8; 0, with a roughly even spread amongst scores of 0, 1, and 2 for this age-range. There is a clearer band of scores at 2 points from ages 9; 0 to 15; 0, potentially indicating more common usage of causally connected actions at around 8–9 years of age.

Finally, the most evident break in scores could be seen for consequence, whereby there was a clear change from the absence of consequence from stories (score of 0) to the presence of consequence at the level of proficiency (score of 2) or elaboration (score of 3) at age 9 and older. As can be seen in the Jitter plot

TABLE 2 Macrostructure total score predicted by age.

	Estimate (β)	Std. Error	t-value	p-value
(Intercept)	0.574	0.456	1.26	0.208
Age	0.972	0.05	19.52	<0.001***

Statistical significance is indicated by *** = < 0.001 ; $R^2 = 0.355$.

TABLE 3 Results of OLR for macrostructure element scores by age.

Model	Estimate (SE)	Odds-ratio [CI]	Wald	p-value
Character ~ Age	0.40 (0.03)	1.49 [1.39, 1.60]	11.3	<0.001***
Setting ~ Age	0.28 (0.03)	1.33 [1.24, 1.42]	8.51	<0.001***
IE ~ Age	0.41 (0.03)	1.50 [1.41, 1.60]	12.74	<0.001***
IR ~ Age	0.12 (0.03)	1.13 [1.07, 1.19]	4.52	<0.001***
Plan ~ Age	0.22 (0.03)	1.25 [1.18, 1.33]	7.6	<0.001***
Action ~ Age	0.35 (0.03)	1.42 [1.33, 1.51]	11.01	<0.001***
Con ~ Age	0.47 (0.03)	1.60 [1.50, 1.71]	13.97	<0.001***

Statistical significance is indicated by *** = < 0.001 . IE, Initiating Event; IR, Internal Response; Con, Consequence.

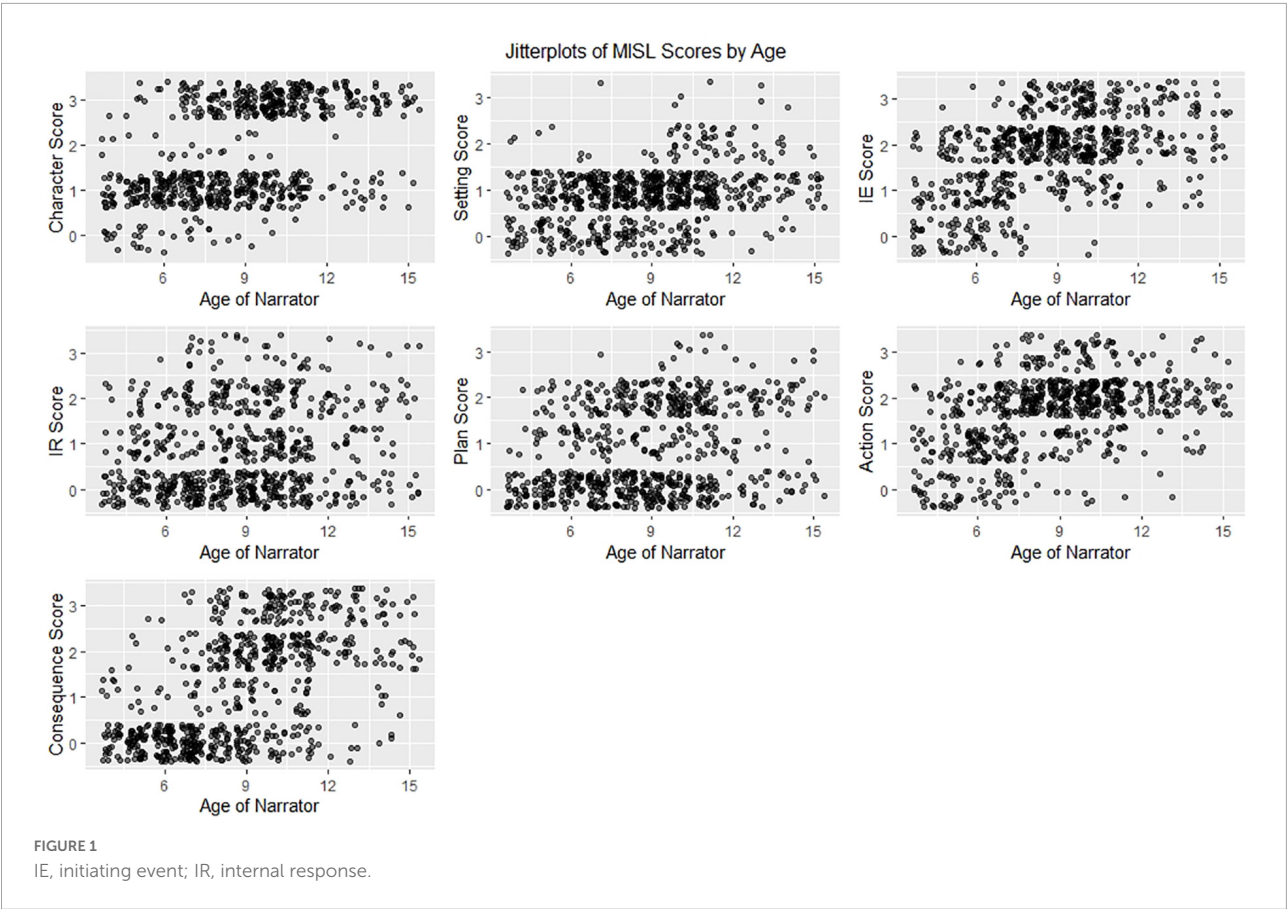


TABLE 4 Modal scores for macrostructure element by age.

	Age of narrator											
	4	5	6	7	8	9	10	11	12	13	14	15
Char	1	1	1	1	1	1	3	3	3	3	3	3
Sett	0	1	1	1	1	1	1	1	1	1	1	1
IE	0	2	1	2	2	2	2	2	2	3	3	2
IR	0	0	0	0	0	0	0	0	2	0	0	2
Plan	0	0	0	0	0	0	2	2	2	2	2	2
Act	0	0	1	2	2	2	2	2	2	2	2	2
Con	0	0	0	0	0	2	2	2	3	3	3	2

Cells highlighted in gray represent where a score ≥ 2 are consistent across increasing ages, indicating proficiency (2) or mastery (3).

for consequence (see Figure 1), the cluster of scores indicates a positive trend whereby older children received higher scores on consequence than younger children.

Discussion

The purpose of this study was to understand the nature of oral narrative macrostructure development in school-age children using a metric based in discourse theory that defines

macrostructure as the story elements and the causal connections between them. This is a departure from prior work based in schema theory that has quantified oral macrostructure abilities by noting the presence or absence of story elements or making holistic judgments about the quality and developmental level of an oral narrative. The MISL incorporates discrete criteria for measuring story elements and their causal connections which removes some of the subjectivity of these earlier rubrics. Our first aim was to determine whether the MISL was sensitive to differences in oral macrostructure abilities across age using

this newer approach. Results indicated that there was strong convergent validity between the TNL-2 Aliens subtest score and the MISL score, $r(686) = 0.766$, $p < 0.001$. Results of both the simple regression on total macrostructure score and the series of ordinal regression analyses for each macrostructure element indicated that age was a significant predictor of the scores children received. Collectively, these results suggest that the MISL is a developmentally valid measure of oral narrative production abilities.

The second research question asked at what age the majority of children in the sample used each macrostructure element in oral narratives. Earlier studies used simple counts and holistic judgments of story elements to measure the quality or developmental level of narratives. The MISL requires that causal connections be explicitly stated in order for specific story elements related to the creation of a complete or complex episode (initiating event, internal response, plan, attempt, consequence) to be given a score of 2 or higher. Further, scores of 2 for character and setting are not given unless 2 or more examples of these story elements are stated in a story. This imposes a more rigid requirement on the scoring of episodes than those based in schema theory. Therefore, we expected to find that our developmental trajectory for the use of macrostructure elements might reflect a lengthier timeline than earlier studies. Following schema theory, Stein and Glenn (1979) asserted that preschool-age children reach the developmental milestone of telling stories that describe characters and list actions chronologically. These findings were supported by Berman and Nir-Sagiv (2007). Consistent with this research, the preschoolers who participated in our study were observed to include characters in their stories. For example, 4-year-olds demonstrated a modal score of 1 for character, and 0 for all other SGEs ($n = 32$).

It was not until age 6 in our sample that we observed higher modal scores of 2 for character, setting and action ($n = 64$) at which time children were describing characters by name and clearly attributing the actions they described in their stories to the characters they introduced. Our observation with regard to this finding is that schema frameworks that do not discreetly measure causality between story elements may be associated with “earlier” achievements in the use of story elements than those based in discourse theory. Similarly, Stein and Glenn (1979) asserted that by age six children typically tell stories that include the aims or intentions of their characters (plans). This was supported by John et al. (2003) where children in their sample were including internal responses and aims of the characters in story retells by the time they were 7-years-old. Using the more rigid criteria imposed by characterization of macrostructure as “including causal connections” our 6-year-olds were not observed to include clearly aligned plans and intentions in their stories, with most scoring a 0 or 1. Scores of 1 would indicate a “planning word” was used (e.g., thought, decided) but without a clear causal relationship to the character, it would not be given

a score of 2 which was required to meet our definition of whether the story element was “present.” The achievement of “complete episode” was reported by Stein and Glenn (1979) and Berman and Nir-Sagiv (2007) between the ages of 7–8 in which children were reported to tell stories that included an initiating event, action, and consequence. Our findings were that the three critical elements defining the achievement of complete episode (initiating event, action, consequence) occurred at 9 years of age. The MISL rubric requires that all three elements (i.e., initiating event, action, consequence) be clearly and specifically connected to each other using specific language (e.g., because, so). Following these criteria, the emergence and stabilization of consequence was most impacted. Initiating event and action stabilized slightly earlier in the current data sample. Our effect size estimates indicated that for each 1-year increase in age, the score for the use of a specific element increased. Thus, as children aged, they were increasingly better at using language to link story elements together using causal language. For example, younger children (i.e., 4 and 5-year-olds) produced stories like, “The car was crashing. The people were walking by the car.” Stories like this received scores of 1 for initiating event and action because a possible initiating event (e.g., The car was crashing) was stated, and an action was stated (e.g., The people were walking by the car.), however, the events were not causally and temporally related. Whereas children who were 9 years of age produced stories like:

The car was about to crash into the big hole, so the people inside started to scream. Then, they pressed on the brakes and turned the wheel to get away from the hole. They missed the hole, and everyone was safe.

Stories like this received a score of 2 for initiating event, action, and consequence—indicating that the three critical elements of a story were included, and they were causally related. In contrast, using a rubric based in schema theory, younger children would receive the same scores as children who were 9 years old in our sample because they would receive scores that reflected whether a story grammar element was present or not.

It was not until age 10 that we observed scores of 2 for the story elements of internal response and plan ($n = 273$). Stories included words that might be associated with planning or feelings the characters may have had, however, students were not observed to consistently use causal language to connect them to the basic episode until much later than reported in earlier studies (Stein and Glenn, 1979; John et al., 2003; Berman and Nir-Sagiv, 2007). It is well supported that individuals from a very young age regularly pay attention to goal motivated actions, plans, and internal responses in the stories they hear or read and tend to include those elements in the stories they create on their own (Lynch and van den Broek, 2007). However, children in the current sample were not shown to consistently use literate language features to establish causal connections between an initiating event and a plan until the age of 10. A child in our sample who was 10 years of age might produce a story like:

The car was about to crash into the big hole. The people inside of the car were scared. Then, they decided to press on the brakes and turned the wheel to try and get away from the hole and that is what they did. They missed the hole, and everyone was safe.

This story would have received a score of 2 for plan because it was temporally related to the initiating event. A score of 1 would have been rewarded for internal response because a feeling was stated. The scores using the MISL for children this age reflect the emergence/presence of some of the story grammar elements (e.g., internal response), and mastery of others (e.g., initiating event, action, plan, consequence). In contrast, using a rubric based in schema theory, the children would have received scores that reflected mastery of all of the story grammar elements by this age because the child included at least one example of the element in their story.

Finally, the emergence of complex episodes in oral narratives was reported by [Stein and Glenn \(1979\)](#) asserted that by the time a child is 11 years of age produce stories that are intricate and include an elaboration of the complete episode. This also has been supported by a variety of studies looking at the complexity of narratives in written contexts ([Dockrell and Connelly, 2016](#); [Jagaiah et al., 2020](#)). Remember that elaboration occurs when a child includes multiple episodes with and more than one plan, action sequence. This was supported in oral narratives by the work of [John et al. \(2003\)](#), where 11-year-old children were found to include elements like internal response at a higher rate than their younger peers. In our sample, it was only children 13 years of age and older ($n = 67$) that were shown to consistently elaborate on their story and include complex episodes in oral narratives. A child who was 13 in our sample might produce a story like:

The white Jeep was about to crash into the big hole in the desert. The people inside of the car were scared. Then, the driver John said, “Hey. Everyone stop screaming so I can think.” He decided to press on the brakes and turned the wheel to try and get away from the hole and that is what they did. They missed the hole, and everyone was safe. Then, all of a sudden, a huge thunderstorm came, and rain started falling fast. Everyone was getting wet, so they decided to drive and find a rock to hide under. That’s what they did. They found a rock and waited until the thunderstorm ended to go home.

This story would have received scores of a 3 for initiating event, plan, action, and consequence because the child included more than one complete story episode where these elements were causally and temporally related. In contrast, using a rubric based in schema theory, a child that produced a more complex story would have received scores similar to those observed at ages 9 or 10 because they would have only received a point based off of the presence/absence of the story grammar elements.

It was not until the age of 15 that we observed scores that reflected “mastery” for the use of internal response (feelings). A child that was 15 years of age may have produced a story like:

The white Jeep was about to crash into the big hole in the desert. The people inside of the car were scared so they started to scream and panic. Then, the driver John said, “Hey. Everyone stop screaming so I can think.” He decided to press on the brakes and turned the wheel to try and get away from the hole and that is what they did. They missed the hole, and everyone was safe. John felt relieved.

A child who produced this story would have received a score of 2 for internal response because the relationship to the feelings was explicitly stated and related to the initiating event. Prior research that has reported the earlier use of internal response at the age of 9 was conducted using story retell data ([Berman and Slobin, 1994](#)). Research has demonstrated that having an adult model in a story retell task has benefited the narrative performance of typically developing children for sentence complexity and story macrostructure ([Sheng et al., 2020](#)). In addition, research has demonstrated that both monolingual and bilingual children include more content in their stories when retelling a story vs. telling a unique story from a picture ([Schneider and Dube, 2005](#); [Lucero and Uchikoshi, 2019](#)). The current research utilized story tells which may require more sophisticated language ability. This may have contributed to the findings that the mastery of this element was not found until the children were 15 years old. It could be that the nature of the task (i.e., creating a story) made it more difficult for the children to utilize complex language to create temporal and causal connections in their stories related to the use of internal response.

Clinical implications

A child’s ability to successfully produce a narrative is an important developmental milestone for school-age children ([Hughes et al., 1997](#)). As narratives are complex in nature, they can be used as a measure of language ability throughout development ([Hudson and Shapiro, 1991](#); [Hughes et al., 1997](#); [National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010](#); [Ukrainetz, 2015](#); [Petersen et al., 2020](#)). Many studies have discussed the usefulness of a schema theory-based approaches which employ measurement of SGEs to examine narrative ability ([Stein and Glenn, 1979](#); [Merritt and Liles, 1987](#); [Berman and Nir-Sagiv, 2007](#); [Bitetti and Hammer, 2021](#)). Not surprisingly, many narrative interventions have been designed that focus on the explicit teaching of SGEs to students who are delayed in their narrative language abilities (for reviews with examples see [Petersen, 2011](#); [Favot et al., 2021](#); [Pico et al., 2021](#)). However, measuring narrative discourse abilities using rubrics based in schema theory may not provide clinicians with a complete picture of the underlying language abilities a child has to bring to the “narrative production table.” This has the potential to result in the use of narrative interventions that do not address the

nature of the difficulties children may experience in becoming proficient in narrative comprehension and production (oral and written). Studies of written discourse have consistently shown that statements in stories that have a large number of causal connections tend to be judged more important, recalled more frequently and retrieved more quickly, than stories with fewer causal connections (Trabasso and Sperry, 1985; O'Brien and Meyers, 1987; Espin et al., 2007). These findings have also been reported for oral discourse (Cevasco and van den Broek, 2008). Current rubrics and narrative macrostructure scoring systems that focus almost entirely on the presence or absence of story elements while asking the rater to make a holistic judgment about whether the story is also “cohesive” in nature may not capture this important aspect of narrative ability.

Limitations and future directions

There were several limitations of our study that are important to consider in the interpretation of the findings. One limitation is the differences in sample sizes at different ages. Participants in this study were drawn from the normative sample for the TNL-2, meaning that we had a larger number of children toward the middle of the age distribution than on the edges of the distribution (i.e., the youngest and oldest ages in the sample). However, because age was normally distributed, we found it appropriate to conduct an ordinal logistic regression to analyze our data and account for differences at each score level by age. Additionally, a potential limitation is found in the generalizability of our findings. The original data used in our analyses came from participants in a few different locations in the United States. Given that the majority of the participants were Caucasian, it is difficult to determine if our results would generalize to children of other ethnic and cultural backgrounds. The benefit of narrative sample analysis, however, is that their use tends to be more sensitive to such differences in backgrounds of participants than standardized assessments (MacLachlan and Chapman, 1988; Dollaghan et al., 1990; Leadholm and Miller, 1992; Wagner et al., 2000; Nippold et al., 2014). Still, additional analyses on a more diverse population of children are needed to better generalize these results to the population of school-age children. As culturally and ethnically diverse populations grow in the United States, it would be beneficial to understand whether results for narrative production would vary across diverse backgrounds.

Finally, stories for this study were from the Aliens subtest from the TNL-2. This study has evidence for the use of this rubric for a spontaneous story-generation prompt. It may be necessary to conduct studies to understand the use of the MISL on stories produced from different elicitation contexts to continue to explore its validity. In the future, it would be beneficial to explore the validity of the MISL across different narrative elicitation contexts. Results might

differ for contexts such as story-retell or personal narratives, which are also commonly used in assessment. In addition, it may be important to understand the use of the MISL with children of differing language abilities, including those who are at-risk for language impairment. This would increase our ability to understand differences in narrative production abilities of typically developing children and those who are at-risk. That knowledge may lead to stronger evidence for the use of interventions targeted at increasing narrative production abilities of children who are at-risk for language impairment.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Internal Review Board at Utah State University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

MI-A contributed to the conception and design of this study and wrote the first draft of this manuscript. CF was main data analyst on project and wrote sections of the manuscript. SH contributed to edits of the first draft of the manuscript. SG was co-PI on data collected for this project and contributed to editing. RG was co-PI on data collected for this project. All authors will contribute to revisions of the submitted version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Applebee, A. N. (1978). *The Child's Concept of a Story: Ages two to Seventeen*. Chicago: University of Chicago Press.
- Berman, R. A. (1988). On the ability to relate events in narrative. *Discourse Process*. 11, 469–497. doi: 10.1080/01638538809544714
- Berman, R. A., and Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Process*. 43, 70–120. doi: 10.1080/01638530709336894
- Berman, R. A., and Slobin, D. I. (1994). “Narrative Structure,” in *Relating Events in Narrative: A Crosslinguistic Developmental Study*, eds R. A. Berman and D. I. Slobin (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc), 39–84.
- Bitetti, D., and Hammer, C. S. (2021). English narrative macrostructure development of Spanish–English bilingual children from preschool to first grade. *Am. J. Speech Lang. Pathol.* 30, 1100–1115. doi: 10.1044/2021_AJSLP-20-00046
- Botting, N. (2002). Narrative as a clinical tool for assessment of linguistic and pragmatic impairments. *Child Lang. Teach. Therapy* 18, 1–22. doi: 10.1191/0265659002ct2240a
- Boudreau, D. M., and Hedberg, N. L. (1999). A comparison of early literacy skills in children with specific language impairment and their typically developing peers. *Am. J. Speech Lang. Pathol.* 8, 249–260. doi: 10.1044/1058-0360.0803.249
- Carvalho, L., Limpo, T., and Pereira, L. A. (2021). The contribution of word-, sentence-, and discourse-level abilities on writing performance: A 3-year longitudinal study. *Front. Psychol.* 12:668139. doi: 10.3389/fpsyg.2021.668139
- Catts, H. W., Fey, M. E., Tomblin, J. B., and Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *J. Speech Lang. Hear. Res.* 45, 1142–1157. doi: 10.1044/1092-4388(2002)093
- Cevasco, J., and van den Broek, P. (2008). The importance of causal connections in the comprehension of spontaneous spoken discourse. *Psicothema* 20, 801–806.
- Crais, E. R., and Lorch, N. (1994). Oral narratives in school-age children. *Top. Lang. Disord.* 14, 13–28. doi: 10.1097/00011363-199405000-00004
- Dockrell, J. E., and Connelly, V. (2016). “The relationships between oral and written sentence generation in English speaking children: The role of language and literacy skills,” in *Written and Spoken Language Development Across the Lifespan*, eds J. Perera, M. Aparici, E. Rosado, and N. Salas (Cham: Springer), 161–177. doi: 10.1007/978-3-319-21136-7_11
- Dockrell, J. E., Connelly, V., Walter, K., and Critten, S. (2014). Assessing children's writing products: The role of curriculum-based measures. *Br. Educ. Res. J.* 41, 575–595. doi: 10.1002/berj.3162
- Dollaghan, C. A., Campbell, T. F., and Tomlin, R. (1990). Video narration as a language sampling context. *J. Speech Hear. Disord.* 55, 582–590. doi: 10.1044/jshd.5503.582
- Espin, C. A., Cevasco, J., van den Broek, P., Baker, S., and Gersten, R. (2007). History as narrative: The nature and quality of historical understanding for students with LD. *J. Learn. Disabil.* 40, 174–182. doi: 10.1177/00222194070400020801
- Favot, K., Carter, M., and Stephenson, J. (2021). The effects of oral narrative intervention on the narratives of children with language disorder: A systematic literature review. *J. Dev. Physical. Disabil.* 33, 489–536. doi: 10.1007/s10882-020-09763-9
- Gillam, R. B., and Pearson, N. A. (2017). *TNL-2: Test of Narrative Language – Second Edition*. Austin: TX: Pro-ed.
- Gillam, S. G., Gillam, R. B., Fargo, J. D., Olszewski, A., and Segura, H. (2017). Monitoring indicators of scholarly language: A progress-monitoring instrument for measuring narrative discourse skills. *Commun. Disord. Quarterly* 38, 96–106. doi: 10.1177/1525740116651442
- Graesser, A. C., Millis, K. K., and Zwaan, R. (1997). Discourse comprehension. *Ann. Rev. Psychol.* 48, 163–189. doi: 10.1146/annurev.psych.48.1.163
- Graham, S., Gillespie, A., and McKeown, D. (2013). Writing: Importance, development, and instruction. *Read. Writ.* 26, 1–15. doi: 10.1007/s11145-012-9395-2
- Greenhalgh, K. S., and Strong, C. J. (2001). Literate language features in spoken narratives of children with typical language and children with language impairments. *Lang. Speech Hear. Ser. Sch.* 32, 114–125. doi: 10.1044/0161-1461(2001)010
- Hedberg, N. L., and Westby, C. E. (1993). *Analyzing Storytelling Skills: Theory to Practice*. Tucson, AZ: Communication Skill Builders.
- Heilmann, J., Miller, J. F., Nockerts, A., and Dunaway, C. (2010). Properties of the narrative scoring scheme using narrative retells in young school-age children. *Am. J. Speech Lang. Pathol.* 19, 154–166. doi: 10.1044/1058-0360(2009)08-0024
- Hudson, J. A., and Shapiro, L. R. (1991). “From knowing to telling: The development of children's scripts, stories, and personal narratives,” in *Developing Narrative Structure*, eds A. McCabe and C. Peterson (Mahwah, NY: Lawrence Erlbaum Associates, Inc), 89–136.
- Hughes, D., McGillivray, L., and Schimidek, M. (1997). *Guide to Narrative Language: Procedures for Assessment*. Eau Claire: Thinking Publications.
- Jagaiah, T., Olinghouse, N. G., and Kearns, D. M. (2020). Syntactic complexity measures: Variation by genre, grade-level, students' writing abilities, and writing quality. *Read. Writ.* 33, 2577–2638. doi: 10.1007/s11145-020-10057-x
- John, S. F., Lui, M., and Tannock, R. (2003). Children's story retelling and comprehension using a new narrative resource. *Can. J. Sch. Psychol.* 18, 91–113. doi: 10.1177/082957350301800105
- Justice, J. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L., and Gillam, R. B. (2006). The index of narrative microstructure: A clinical tool for analyzing school-age children's narrative performances. *Am. J. Speech Lang. Pathol.* 15, 177–191. doi: 10.1044/1058-0360(2006)017
- Kintsch, W., and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychol. Rev.* 85, 363–394. doi: 10.1037/0033-295X.85.5.363
- Leadholm, B. J., and Miller, J. F. (1992). *Language Sample Analysis: The Wisconsin Guide*. Wisconsin: Wisconsin Department of Public Instruction.
- Liles, B. (1993). Narrative discourse in children with language disorders and children with normal language. *J. Speech Lang. Hear. Res.* 36, 868–882. doi: 10.1044/jshr.3605.868
- Liles, B., Duffy, R. J., Merritt, D. D., and Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *J. Speech Lang. Hear. Res.* 38, 415–425. doi: 10.1044/jshr.3802.415
- Lucero, A., and Uchikoshi, Y. (2019). Narrative assessments with first grade Spanish-English emergent bilinguals: Spontaneous versus retell conditions. *Narrat. Inq.* 29, 137–156. doi: 10.1075/ni.18015.luc
- Lynch, J. S., and van den Broek, P. (2007). Understanding the glue of narrative structure: Children's on- and off-line inferences about character's goals. *Cogn. Dev.* 22, 323–340. doi: 10.1016/j.cogdev.2007.02.002
- Mackie, J. L. (1980). *The Cement of the Universe: A Study of Causation*. Oxford: Oxford (Clarendon) University Press. doi: 10.1093/0198246420.001.0001
- MacLachlan, B. G., and Chapman, R. S. (1988). Communication breakdowns in normal and language learning-disabled children's conversation and narration. *J. Speech Hear. Disord.* 53, 2–7. doi: 10.1044/jshd.5301.02
- Mandler, J. M., and Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cogn. Psychol.* 9, 111–151. doi: 10.1016/0010-0285(77)90006-8
- Mayer, M. (1967). *A Boy, a Dog and a Frog*. New York, NY: Dial Press.
- Mayer, M. (1969). *Frog. Where Are You?*. New York, NY: Dial Press.
- Mayer, M. (1974). *Frog Goes to Dinner*. New York, NY: Dial Press Books for Young Readers.
- Mayer, M. (1975). *One Frog too Many*. New York, NY: Dial Press Books for Young Readers.
- Merritt, D. D., and Liles, B. Z. (1987). Story grammar ability in children with and without language disorder: Story generation, story retelling, and story comprehension. *J. Speech Lang. Hear. Res.* 30, 539–552. doi: 10.1044/jshr.3004.539
- Meyer, B. J. F. (1975). *The Organization of Prose and its Effects on Memory*. Amsterdam: North-Holland Publishing Company.
- Miles, S., and Chapman, R. S. (2002). Narrative content as described by individuals with Down syndrome and typically developing children. *J. Speech Lang. Hear. Res.* 45, 175–189. doi: 10.1044/1092-4388(2002)013
- Miller, J., Andriacchi, K., DiVall-Rayn, J., and Lien, P. (2003). *Narrative Scoring Scheme*. Madison, MDN: SALT Software, LLC.
- Miller, J., and Iglesias, A. (2019). *Systematic Analysis of Language Transcripts (SALT), Research Version 20 [Computer Software]*. Madison, MDN: SALT Software, LLC.
- Munoz, M. L., Gillam, R. B., Pena, E. D., and Gulley-Faehnle, A. (2003). Measures of language development in fictional narratives of Latino children. *Lang. Speech Hear. Ser. Sch.* 34, 332–342. doi: 10.1044/0161-1461(2003)027

- National Governors Association Center for Best Practices and Council of Chief State School Officers. (2010). *Common Core State Standards for English Language Arts*. Washington, DC: National Governors Association Center for Best Practices.
- Nicolosi, L., Harryman, E., and Kresheck, J. (2004). *Terminology of communication disorders: Speech-language-hearing*, 5th Edn. Baltimore, MD: Lippincott Williams & Wilkins.
- Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., and MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: Examining the use of complex syntax. *J. Speech Lang. Hear. Res.* 57, 876–886. doi: 10.1044/1092-4388(2013)13-0097
- O'Brien, E. J., and Meyers, J. L. (1987). The role of causal connections in the retrieval of text. *Memory Cogn.* 15, 419–427. doi: 10.3758/BF03197731
- Petersen, D. B. (2011). A systematic review of narrative-based language intervention with children who have language impairment. *Commun. Disord. Quarterly* 32, 207–220. doi: 10.1177/1525740109353937
- Petersen, D. B., Spencer, T. D., Konishi, A., Sellars, T. P., Foster, M. E., and Robertson, D. (2020). Using parallel, narrative-based measures to examine the relationship between listening and reading comprehension: A pilot study. *Lang. Speech Hear. Ser. Sch.* 51, 1097–1111. doi: 10.1044/2020_LSHSS-19-00036
- Peterson, C., and McCabe, A. (1983). *Developmental Psycholinguistics: Three Ways of Looking at a Child's Narrative*. New York, NY: Plenum Press. doi: 10.1007/978-1-4757-0608-6
- Pico, D. L., Hessling, Pahl, A., Biel, C. H., Peterson, A. K., Biel, E. J., Woods, C., et al. (2021). Interventions designed to improve narrative language in school-age children: A systematic review with meta-analyses. *Lang. Speech Hear. Ser. Sch.* 52, 1109–1126. doi: 10.1044/2021_LSHSS-20-00160
- Purcell, S. L., and Liles, B. Z. (1992). Cohesion repairs in the narratives of normal-language and language-disordered school-age children. *J. Speech Lang. Hear. Res.* 35, 354–362. doi: 10.1044/jshr.3502.354
- Reilly, J., Losh, M., Bellugi, U., and Wulfeck, B. (2004). Frog, where are you? Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain Lang.* 88, 229–247. doi: 10.1016/S0093-934X(03)00101-9
- Roth, F. P., and Speckman, N. J. (1986). Narrative discourse: Spontaneously generated stories of learning-disabled and normally achieving students. *J. Speech Lang. Hear. Pathol.* 51, 8–23. doi: 10.1044/jshd.5101.08
- Roth, F. P., Speech, D. L., and Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *J. Educ. Res.* 95, 259–272. doi: 10.1080/00220670209596600
- Rumelhart, D. E. (1975). "The active structural network," in *Explorations in cognition*, eds D. A. Norman and D. E. Rumelhart (San Francisco, CA: W. H. Freeman).
- Rumelhart, D. E. (1977). *Introduction to Human Information Processing*. New York, NY: Wiley.
- Schneider, P., and Dube, R. V. (2005). Story presentation effects on children's retell content. *Am. J. Speech Lang. Pathol.* 14, 52–60. doi: 10.1044/1058-0360(2005/007)
- Sheng, L., Shi, G., Wang, D., Hao, Y., and Zheng, L. (2020). Narrative production in Mandarin-speaking children: Effects of language ability and elicitation method. *J. Speech Lang. Hear. Res.* 63, 774–792. doi: 10.1044/2019_JSLHR-19-00087
- Stadler, M. A., and Ward, G. C. (2005). Supporting the narrative development of young children. *Early Child. Educ. J.* 33, 73–80. doi: 10.1007/s10643-005-0024-4
- Stein, M. L., and Glenn, C. (1979). "Analysis of story comprehension in elementary school children," in *New Directions in Discourse Processing*, Vol. 2, ed. R. O. Freedle (New York, NY: Ablex), 53–120.
- Stein, N. L. (1988). "The development of children's storytelling skill," in *Child Language: A Reader*, eds M. B. Franklin and S. Barten (Oxford: Oxford University Press), 282–297.
- Strong, E. J. (1998). *The Strong Narrative Assessment Procedure*. Texas: Thinking Publications.
- Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cogn. Psychol.* 9, 77–110. doi: 10.1016/0010-0285(77)90005-6
- Trabasso, T., and Sperry, L. L. (1985). Causal relatedness and importance of story events. *J. Memory Lang.* 24, 595–611. doi: 10.1016/0749-596X(85)90048-8
- Ukrainetz, T. A. (2015). "Telling a good story: Teaching the structure of narrative," in *School-Age Language Intervention: Evidence-Based Practices*, ed. T. A. Ukrainetz (Austin: Pro-ed), 335–378.
- van den Broek, P. (1990). "Causal inferences and the comprehension of narrative text," in *Inferences and text Comprehension*, eds A. C. Graesser and G. H. Bower (Cambridge, MA: Academic Press), 175–196. doi: 10.1016/S0079-7421(08)60255-8
- van Dijk, T. A., and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Cambridge, MA: Academic.
- Wagner, C. R., Nettelbladt, U., Sahlén, B., and Nilholm, C. (2000). Conversation versus narration in pre-school children with language impairment. *Int. J. Lang. Commun. Disord.* 35, 83–93. doi: 10.1080/136828200247269
- Westerveld, M. F., Gillon, G. T., and Miller, J. F. (2004). Spoken language samples of New Zealand children in conversation and narration. *Adv. Speech Lang. Pathol.* 6, 195–208. doi: 10.1080/14417040400010140
- Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychol. Sci.* 6, 292–297. doi: 10.1111/j.1467-9280.1995.tb00513.x



OPEN ACCESS

EDITED BY

David Scheer,
Ludwigsburg University of Education,
Germany

REVIEWED BY

Christine Espin,
Leiden University, Netherlands
Dennis Christian Hövel,
Interkantonale Hochschule für
Heilpädagogik (HfH), Switzerland

*CORRESPONDENCE

Alina Hase
hase@leuphana.de

SPECIALTY SECTION

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

RECEIVED 14 April 2022

ACCEPTED 29 July 2022

PUBLISHED 18 August 2022

CITATION

Hase A, Kahnbach L, Kuhl P and Lehr D
(2022) To use or not to use learning
data: A survey study to explain German
primary school teachers' usage of data
from digital learning platforms
for purposes of individualization.
Front. Educ. 7:920498.
doi: 10.3389/feduc.2022.920498

COPYRIGHT

© 2022 Hase, Kahnbach, Kuhl and
Lehr. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

To use or not to use learning data: A survey study to explain German primary school teachers' usage of data from digital learning platforms for purposes of individualization

Alina Hase^{1*}, Leonie Kahnbach¹, Poldi Kuhl^{1,2} and Dirk Lehr^{1,3}

¹Competencies for Digitally-Enhanced Individualized Practice (CODIP), Department of Education, Leuphana University Lüneburg, Lüneburg, Germany, ²Department of Education, Institute of Educational Research, Leuphana University Lüneburg, Lüneburg, Germany, ³Department of Education, Institute of Psychology, Leuphana University Lüneburg, Lüneburg, Germany

Digital learning platforms (DLP) provide various types of information about student learning when used for learning and practice. This learning data holds potential for individualized instruction, which has become increasingly necessary for adequately addressing learners' individual needs. For primary schools in particular, this is important for developing inclusive schools. However, despite the potential of DLP and the learning data that can be obtained from them, they are rarely used by teachers. Furthermore, little is known about factors that lead teachers to use learning data for instruction and individual support. To address this research gap, we conducted an online cross-sectional survey study of $N = 272$ primary school teachers in Germany. After describing the respondents' current and previous usage of learning data from DLP, we used structural equation modeling (SEM) to test the influence of predictors on respondents' intention to use as well as their usage of learning data from DLP. Finally, we discuss the need for increased usage of learning data in teacher education and training, contributing to ongoing debates about the usage of digital learning data in educational research and practice.

KEYWORDS

Data-Based Decision Making, digital learning platforms, individualization, Learning Analytics, primary school teacher, structural equation modeling, Theory of Planned Behavior

Introduction

Today, teachers face a variety of challenges in their daily school life, including an increased number of administrative tasks, a heterogeneous student population, and the digitization of schools (Schmid et al., 2017; Tondeur et al., 2018). Overcoming these challenges can be exhausting, but also holds potential for the further development

of schools and teaching. More precisely, combining different challenges can provide additional opportunities for development. For example, digital learning platforms (DLP) can support the development of individual learning requirements in the context of inclusive education. This is especially relevant for primary schools due to their high heterogeneity (Schwab et al., 2017). The usage of DLP contributes to the ability of all learners to participate in the classroom (Vanbecelaere et al., 2020; Schaumburg, 2021). Regarding the support of learners, DLP hold an added value of particular importance: the availability of learning data. Using this learning data, teachers can track and reflect on individual learning processes and implement appropriate learning support (FitzGerald et al., 2018). To date, such data usage is found primarily in research on Data-Based Decision Making (DBDM) and Learning Analytics (e.g., Mandinach and Schildkamp, 2020; Blumenthal et al., 2021; Krein and Schiefner-Rohs, 2021). Although the benefits of using (digital) learning data have been highlighted in previous research, the factors that promote or hinder primary school teachers' usage of learning data from DLP, especially in Germany, remain mostly unconsidered. Therefore, based on the Theory of Planned Behavior (TPB) (Ajzen, 1991) and in consideration of further potentially influential factors, we conducted a cross-sectional survey study among German primary school teachers to investigate the antecedents of their intention to use and usage of learning data from DLP.

First, we consider digital media which collect learning data for the teacher, DLP, and their potential for individualized practice. We then address the usage of learning data in the context of instructional design. These two topics provide the contextual basis for the study to examine the intention to use and usage of learning data from DLP. Second, the introduction of the TPB will allow us to predict the intention to use as well as the usage of learning data from DLP based on an established model.

Digital learning platforms for individualized practice

According to Böhme et al. (2020), the aims of using digital media in schools are, on the one hand, the promotion of a critical use of digital media and, on the other hand, the support of learning. Especially in highly heterogeneous inclusive school settings, there is a great potential of digital media, as digital media have the potential to increase the participation of all students in the classroom (Vanbecelaere et al., 2020; Schaumburg, 2021). In this context, digital media can support teachers with diagnostic information and thus foster individualized learning offers (Schaumburg, 2021).

There are several types of digital media that can improve student learning, such as intelligent tutorial systems, drill-and-practice programs, or learning management systems

(Nattland and Kerres, 2009; Petko, 2014). However, in the context of this study, we did not study a specific type of digital media. Reinhold et al. (2020, p. 1) emphasize “that it is not the mere medium that does have an effect on learning outcomes, but rather the appropriate way of implementing it into the classroom as well as certain features that technology enhanced learning environments can offer.” Therefore, the focus is on digital media that are used for individualized practice. To specify these, the term DLP is used comprehensively. As examples for DLP Anton or Bettermarks can be mentioned (Holmes et al., 2018; Schaumburg, 2021). Here, DLP contain the following characteristics: First, DLP include practice exercises that students can work on (Greller et al., 2014; Daniela and Rüdolf, 2019). During the assignment, DLP analyze the students' input and provide them with direct, formative feedback (Daniela and Rüdolf, 2019; Hillmayr et al., 2020). At the same time, student results (as an example for learning data) are stored and displayed on a teacher-dashboard within the DLP (Greller and Drachsler, 2012; Greller et al., 2014). Furthermore, DLP create interaction with students and teachers as well as between them (Faustmann et al., 2019; Hillmayr et al., 2020). In the best case, DLP include the possibility of adaptive adjustments by the system itself or the teacher (Daniela and Rüdolf, 2019). In particular, by using learning data provided by DLP, teachers have the opportunity to provide individualized instruction to their students. The usage of learning data for instructional design—with both digital and analog data—has already been addressed in several research fields, which are presented in the following chapter.

Usage of learning data for instructional design

Under the term of DBDM—which refers to “the systematic collection and analysis of different kinds of data to inform educational decisions” (Mandinach and Schildkamp, 2020, p. 1)—learning data in teaching and learning processes became a major focus of research. Among others, DBDM can help teachers determine instructional steps that meet learners' diverse needs (Mandinach and Gummer, 2016; Prenger and Schildkamp, 2018; Peters et al., 2021). Because DBDM focuses on every child, not just children with identified special educational needs, it is consistent with the idea of an inclusive school environment (Mandinach and Gummer, 2013; Knickenberg et al., 2020). DBDM assumes that teachers collect a variety of data (i.e., quantitative, qualitative; analog, digital) in their daily practice. However, as a large amount of data is available, especially in the age of digitalization, teachers should consider which data they want to use and for what purpose (Schildkamp, 2019). Since the usage of DLP should be embedded in a pedagogical concept, the usage of learning data and the resulting decision making evokes

pedagogical actions (Molenaar and Knoop-van Campen, 2017; Kerres, 2018). Therefore, on the one hand, teachers need pedagogical knowledge, perceptions, and an openness to the fact that pedagogy changes with the integration of digital media and learning data. A positive attitude toward the usage of learning data is considered essential (Blumenthal et al., 2021). On the other hand, teachers need data literacy to analyze and appropriately interpret learning data and to set and implement learning goals (Schildkamp and Kuiper, 2010; Mandinach and Gummer, 2016; Molenaar and Knoop-van Campen, 2017; Krein and Schiefner-Rohs, 2021).

In recent years, the research field of Learning Analytics evolved. This can be seen as a further development of research on DBDM. Here, the usage of learning data is considered only in a digital context. Learning Analytics help teachers and learners to individualize learning processes based on digital learning data (Krein and Schiefner-Rohs, 2021). The idea of Learning Analytics grew due to the large amount of learning data collected with the help of digital technologies (Greller et al., 2014). Learning Analytics deal with digitally generated data that is analyzed and presented in real time (Ifenthaler and Drachsler, 2020). Learning Analytics and DBDM pursue the same goal: Both concepts aim to support teachers in making pedagogical decisions based on learning data and not only on experience and intuition, for example, to enable individualized learning processes (Schildkamp and Kuiper, 2010; Greller et al., 2014). Digital learning data, as collected by DLP, include, for example, how long students practiced with the DLP, how many tasks they worked on, and whether they solved the tasks correctly or incorrectly.

Despite the potential of learning data for instructional design, it is, yet, not being used to a great extent for decision making among teachers (Schildkamp and Kuiper, 2010; Kippers et al., 2018). Especially in Germany, few teachers and schools have practiced DBDM to date. However, in other countries such as the United States and the Netherlands, DBDM is already being implemented more frequently (Blumenthal et al., 2021;

Schaumburg, 2021). Studies of DBDM have found positive effects for students and teachers in primary schools. For example, Keuning et al. (2019) showed that teachers' data usage had a positive impact on student achievement in mathematics and spelling. Anderson et al. (2020) found that progress monitoring—which is also a DBDM concept to identify learning problems—can help students acquire reading skills and help teachers to address student heterogeneity. Further, Souvignier et al. (2021) reported that student achievement in reading and mathematics improved after a progress monitoring intervention. Peters et al. (2021) identified the potential of DBDM for teachers dealing with particularly low-performing students. Molenaar and Knoop-van Campen (2018) observed teachers' usage of learning data from DLP and found that Dutch teachers referred to the learning data multiple times during their instruction and that the data influenced their pedagogical actions. Although some studies have addressed DBDM and Learning Analytics in school contexts, there is still a need for further research. The usage of learning data retrieved especially from DLP and the intention to use the learning data for individualized instruction has not been covered empirically. Therefore, a need for research is indicated (Molenaar and Knoop-van Campen, 2017; Blumenthal et al., 2021; Schaumburg, 2021).

Explaining teachers' behavioral intention and usage of learning data

In order to better understand why primary school teachers use data from DLP (or not), further studies are needed. To gain insights into the factors which are associated to teachers' intention to use and their usage of learning data from DLP, we refer to the TPB (Figure 1; Ajzen, 1991) as a theoretical framework. Also, teacher-specific factors are considered additionally. Since this research focuses on the usage of learning data from DLP and not on the usage of the DLP as a technology, the TPB is preferred over

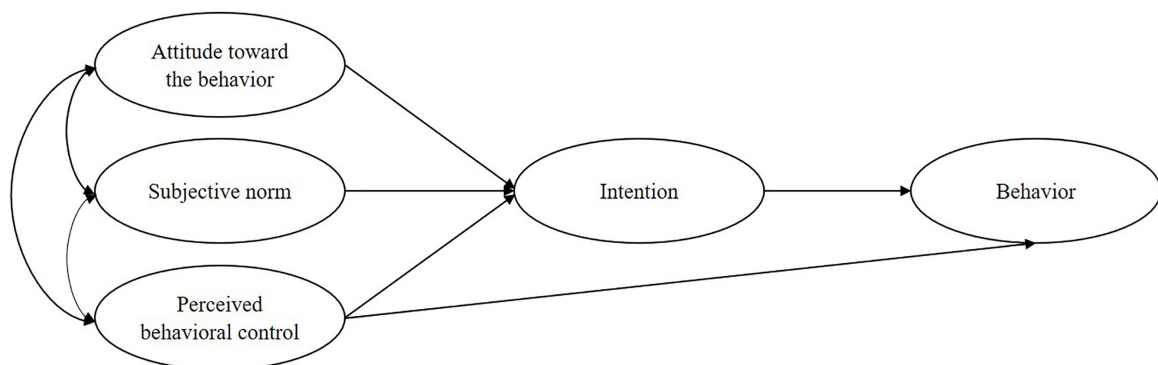


FIGURE 1
The Theory of Planned Behavior [oriented to Ajzen (1991, p. 182)].

models that address acceptance of technology, such as UTAUT (Venkatesh et al., 2003).

The TPB was developed as an extension of the Theory of Reasoned Action (Fishbein and Ajzen, 1975) and aims to explain and predict human behavior in various situations (Ajzen, 1991). Fundamental to the explanation of behavior is that actual behavior is predicted by intended behavior. A high behavioral intention increases the probability for actual behavior. In some instances, there is a gap between intention and behavior. Accordingly, though a person might have a specific intention, the behavior is not always performed (Sheeran, 2002). The TPB attempts to explain this intention-behavior-gap by considering the factors that influence intention as well as by examining the relationship between the intention and the actual behavior (Sheeran, 2002). According to the TPB, an intention is predicted by attitudes toward the behavior, subjective norms, and perceived behavioral control (Ajzen, 1991). Attitude refers to a person's positive or negative feelings toward a behavior. The subjective norm includes the expectations a person has about the reactions of others. Finally, perceived behavioral control contains the estimation of the person's skills, competencies, and resources to perform the behavior (Nistor, 2020).

In educational research, TPB has proven to be a useful instrument to explain teachers' intentions and actual behavior. TPB has been used in (primary) teaching studies on inclusive education and individualized student support (Hellmich et al., 2019; Knauder and Koschmieder, 2019), DBDM (Pierce et al., 2013; Prenger and Schildkamp, 2018), and technology acceptance (Teo and Tan, 2012). In a cross-sectional survey study on $N = 290$ German primary school teachers, Hellmich et al. (2019) identified the school principal's expectations as indicators of subjective norms as the largest factor influencing teachers' intentions to implement inclusive education. Positive attitudes toward inclusion and teachers' self-efficacy beliefs about organizing inclusive education were also, but were found to be less important. Knauder and Koschmieder (2019) applied the TPB to investigate teachers' intrinsic and extrinsic motivational intentions to support students individually as well as to predict teachers' individualized support and lesson design in a cross-sectional survey study involving $N = 488$ Austrian primary school teachers. For the intrinsic motivational intention, they found a strong association between attitude and individualized support, but no significant influence of subjective norms and perceived behavioral control (Knauder and Koschmieder, 2019). This was different for extrinsic motivational intention. Here, school as a factor of subjective norms had a significant influence on the intention to support students, whereas attitude and perceived behavioral control did not predict their extrinsic motivational intention to support students individually (Knauder and Koschmieder, 2019). For the context of DBDM, the TPB model was also able to explain teachers' intention to use data from different sources to inform their teaching. Also, in a cross-sectional survey

study of approximately 1,000 Australian teachers, Pierce et al. (2013) used the TPB model to gain insights into teachers' perceptions of factors influencing their intention to use data from national testing in their lesson planning, confirming the usefulness of the TPB model in explaining teachers' intention to use data. Prenger and Schildkamp (2018) tested an extended version of the TPB model (affective and instrumental attitude, subjective norms, perceived behavioral control, self-efficacy, collective efficacy) to explain teachers' intentions as well as their instructional data usage related to curriculum assessments. They conducted a cross-sectional survey study with $N = 131$ primary school teachers in the Netherlands. Perceived behavioral control predicted instructional data usage, whereas, intention to use data was significantly predicted only by affective attitude and instrumental attitude (Prenger and Schildkamp, 2018). Teo and Tan (2012) reported that TPB is a useful instrument for explaining technology acceptance in educational contexts. In a cross-sectional survey study of $N = 293$ Singapore pre-service teachers, attitude toward technology were found to have the greatest influence on intention to use technology. Perceived behavioral control and subjective norms were also identified but were found to be less important predictors (Teo and Tan, 2012). The abovementioned studies indicate the relevance of the TPB model in educational research. To date, however, no study has used TPB to examine teachers' usage of learning data received from digital media such as DLP in the context of individualization.

Since "teaching is an activity where teachers enact their conceptions about teaching and learning" (Yan et al., 2021, p. 229), it might be useful to consider other factors besides the TPB Model in order to gain more insight. The TPB model includes only a few personal factors and these factors are particularly related to the investigated behavior. However, a systematic review demonstrated the relevance of other factors, such as teaching beliefs, to the implementation of formative assessment (Yan et al., 2021). Accordingly, it can be assumed that the didactic context in which learning data from DLP is used should be considered. Learning data from DLP is related to the usage of DLP, is embedded in practice, contains feedback, is used for individualization, and is to be regarded overall in the context of data-based instructional design. Other studies that have examined the usage of media didactics in the classroom in general hypothesized that didactical concepts have an influence on the intention to use digital media in schools (Tappe, 2017; Gellerstedt et al., 2018). This assumption is also adopted in the present study and will be tested to predict teachers' intention to use learning data obtained from DLP.

Purpose and research questions

The theoretical overview above illustrated that digital media, especially DLP, are suitable to address the challenges of

heterogeneous groups of students and to support and encourage learners individually. Learning data from DLP are of great importance in this context. Research on DBDM and Learning Analytics highlights the utility of learning data for instructional design. However, the reasons why teachers intend to use, use or do not use learning data from DLP have not yet been investigated. Therefore, this study—with reference to the TPB—seeks to answer the following research questions:

1. To what extent do primary school teachers in Germany use learning data from DLP for individualization?
2. What predicts teachers' intentions to use and the usage of learning data from DLP?

Materials and methods

Study design

To gain insights into German primary school teachers' technology and data acceptance, we conducted an online cross-sectional survey study using LimeSurvey from October to December 2021. The study was developed from a psychological and educational perspective as part of the Competencies for Digitally-Enhanced Individualized Practice (CODIP) project. In this article, we focus on clarifying teachers' intentions to use and their usage of learning data from DLP. An examination of the acceptance of DLP itself will be provided in another publication. Primary school teachers from the northern German federal states Bremen, Hamburg, Mecklenburg-Western Pomerania, Lower Saxony, and Schleswig Holstein were recruited by sending emails to their schools. In addition, teachers were reached *via* social media, and we also commissioned a market research panel to recruit teachers. Participation in the online survey took an average of 20 min.

The study was preregistered at Open Science Framework before we accessed the research data.¹ Additionally, because the study involved human subjects, it was reviewed and approved by the Ethics Committee of the Leuphana University Lüneburg. Furthermore, the study was approved by the respective education offices of each involved federal state.

Participants

To find participants, 2,684 schools in northern Germany were contacted. The resulting total sample of the study consisted of $N = 272$ primary school teachers who were predominantly female (86%). This distribution corresponds to the findings obtained by the survey of the Federal Statistical Office for the school year 2020/21 (Federal Statistical Office of Germany, 2022). Most participants

were between 40–49 and 50–59 years old (each 28%), 19% were between 30–39 years, 13% were older than 60 years, and 12% were younger than 30 years. The distribution is roughly in line with the information from the Federal Statistical Office, in which the 40–49-year age group is the largest, with the 30–39-year age group second, followed by the 50–59-year age group (Federal Statistical Office of Germany, 2022). Work experience was also captured by time ranges. Most of the teachers had 11–20 years of work experience (28%), followed by 26% with 21–30 years of work experience, while 20% of the teachers were career starters with up to 5 years of work experience. The remaining teachers had 6–10 years (13%), 31–40 years (11%), and more than 40 years (2%) of work experience.

Survey instrument

The online survey started with an introductory video. This self-created video included a definition of DLP and its resulting learning data. The definitions were intended to ensure that all participants had the same understanding of DLP and learning data from DLP. Similarly, teachers who do not have experience with learning data from DLP can answer the questionnaire based on the video. The questions were divided into four sections: (1) DLP, (2) learning data from DLP, (3) digital media, and (4) didactical concepts. The sections relevant for this article Sections “Purpose and research questions,” Materials and methods,” and “Results” will be described in detail below. All scales used in the survey were adapted from existing measures or were based on theoretical assumptions. The item wordings were altered to the context of usage of data from DLP. Table 1 summarizes all characteristics of the scales used within this research context.

Learning data from digital learning platforms

Following TPB (Ajzen, 1991) as a theoretical framework to explain teachers' intentions to use learning data from DLP, we used scales to assess teachers' self-reported data usage, their intentions to use learning data from DLP, their attitude toward learning data from DLP, their perceived behavioral control regarding learning data from DLP and their subjective norm regarding learning data from DLP.

To assess *data usage* in the context of individualization we developed items on purpose-related usages like using data from DLP for identifying student needs, setting learning goals, or revising lessons based on individual needs. Because to date no study had examined teachers' usage of learning data from DLP, we adapted items on general usage of learning data to this context. This also applies to the following scales concerning learning data from DLP.

Within this questionnaire, *intention to use data from DLP* was measured using three different stages. Thus, the items addressing the intention to use included thinking about a behavior, planning the behavior, as well as the determination of the intention to actually perform the behavior. Still, if teachers answered that they have not used learning data from DLP yet,

¹ <https://doi.org/10.17605/OSF.IO/PG6R4>

TABLE 1 Scale characteristics.

	Number of items	Response scale	Most representative item	References
Usage of data from DLP	5	1 = No, never for this reason–4 = Yes, regularly for this reason	“I use data from DLP to set learning goals for individual students”	Moore and Shaw, 2017
Intention to use data from DLP	3	1 = No–4 = Yes	“I plan to use data from digital learning platforms for my teaching within the current school term”	Venkatesh et al., 2003; Tappe, 2017
Attitude toward data from DLP	8	1 = Does not apply at all–4 = Applies completely; 5 = I cannot tell	“I find data from digital learning platforms useful”	Wayman et al., 2016
Perceived behavioral control regarding data from DLP	4	1 = Does not apply at all–4 = Applies completely; 5 = I cannot tell	“I am good at adapting lessons based on data from digital learning platforms”	Wayman et al., 2016
Subjective norm regarding data from DLP	4	1 = Very low level–4 = Very high level; 5 = I cannot tell	“Colleagues influence me to use data from DLP”	Venkatesh et al., 2003; Tappe, 2017
Usage of digital media	10	1 = Yes 2 = No	“learning videos such as YouTube”	Schmid et al., 2017
Attitude toward digital media	8	1 = Does not apply at all–4 = Applies completely	“I like to use digital media for my lesson planning”	Venkatesh et al., 2003; Tappe, 2017; Petko et al., 2018; Schaumburg and Prasse, 2019
Practice	6	1 = Never–4 = Regular	“I let my students practice with tasks where I can see particularly well whether the essentials have been understood”	Jäger and Helmke, 2008; Baumert et al., 2009
Individual support	4	1 = Never–4 = Regular	“I give the students different tasks depending on their ability”	Institute for Quality Development Hessen, 2012
Feedback	4	1 = Never–4 = Regular	“I tell the students in which areas they can still improve”	PISA, 2017
Data-based instructional design	6	1 = Never–4 = Regular	“I use data as the basis for conversations with parents”	Wayman et al., 2016

All items were taken from the given references and adapted to the context of this study.

they got an additional information before answering the items regarding the intention to use learning data from DLP. We asked teachers to answer the next items to the best of their ability, and to think about potential use if necessary.

Based on theoretical background, *attitude*, *perceived behavioral control*, and *subjective norm* should predict teachers' intention to use learning data from DLP and their usage. Regarding *attitudes toward learning data from DLP*, items were devoted to the benefits of data usage for the teacher, as well as items focused on improvements for students. The items on *perceived behavioral control regarding learning data from DLP* captured how teachers assessed their own ability to use learning data from DLP. Here, using learning data was again focused on aspects of individualization and additionally on aspects of instructional design. Within the items on *subjective norm regarding learning data from DLP*, teachers were asked to rate how much they think other professionally relevant groups of people (students, parents, colleagues, school administrators) expect them to use learning data from DLP. We also provided an option to give no answer as the items of these scales were mandatory to be able to proceed the questionnaire.

Digital media

In addition to the TPB model, research data on other factors were collected to elucidate the intention to use learning data from DLP. For this purpose, we asked teachers about the *digital media they use*. Teachers had to indicate for ten

different digital media whether they use them as part of their teaching. The focus was less on technical devices and more on applications such as learning management systems or learning videos. A sum score was calculated across the ten items to indicate teachers' proneness to usage of digital media for instruction. Additionally, the questionnaire contained items regarding teachers' *attitudes toward digital media*. These items also contained positive attitudinal statements regarding benefits to teachers and students, but with focus on digital media in general.

Didactical concepts

When using DLP and the resulting data, the pedagogical context requires closer consideration. Therefore, we included items on didactical concepts in our survey instrument. Related to our research aim to better understand teachers' intention to use learning data from DLP we integrated items regarding *practice*, *individual support*, *feedback*, and *data-based instructional design* within our study. The didactic concepts were not related to digitalization in order to find out how important these concepts were for teachers in their lessons independently from the usage of data from DLP.

The scale *Practice* consisted of items assessing automated and elaborated practice as those are different ways to practice. Since individual instructional design can benefit from the usage of learning data from DLP, items on *Individual support* were used to its importance for teachers' instruction. The scale

Feedback inquired the extent to which teachers provide feedback to their students in the classroom. With the scale *Data-driven instructional design*, we determined whether teachers also use learning data without a digital medium to design their lessons and to encourage and challenge learners. A mean value was calculated for each scale.

Statistical methods

Missing data

Non-response occurred when participants did not answer all items of a scale or used the alternative response option, *I cannot tell*. To cope with missing data, we used multiple imputation (Van Buuren, 2012), using the mice package in R-Studio. The multiple imputation was based on a created predictor matrix with 10 iterations.

Data analysis

To answer the first research question, the research data on usage of learning data from DLP were analyzed descriptively. Descriptive statistics were compiled for all scales in preparation for explaining the intention to use and the usage of learning data from DLP.

To answer the second research question, we used structural equation modeling (SEM). The SEM method was chosen as it enables to consider all variables in one model at once, acting as both independent and dependent variables. SEM combines factor and path analysis in order to separate measurement error influences from true influences. In addition, SEM can be used to check the fit of a model with the data set (Schumacker and Lomax, 2010; Eid et al., 2017).

We first examined the TPB model as our Model 1. Since this was applied to a novel context, we checked all connections of the variables (i.e., also attitude and subjective norm for usage and not only for intention to use). In another SEM, Model

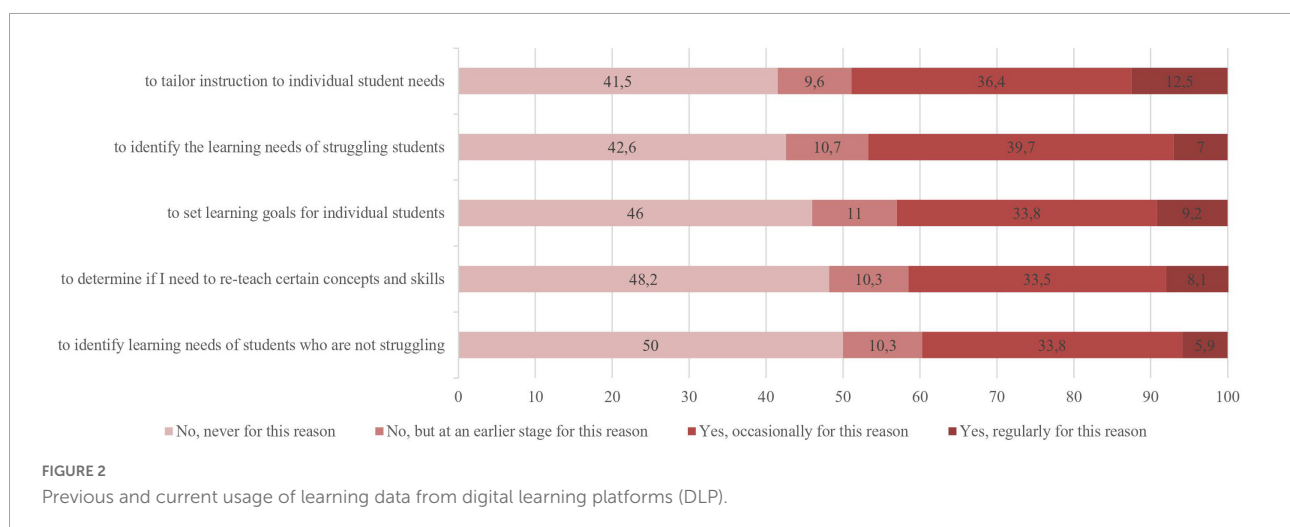
2, we tested the extended TPB model that included additional predictors. Following Tappe (2017) and Gellerstedt et al. (2018), the additional predictors are tested only in relation to the intention to use learning data from DLP and not in relation to the self-reported usage. We specified that the difference in the explained variance between both models (ΔR^2) must be ≥ 0.05 for the extended model to provide a meaningful improvement compared to Model 1. For the purpose of this article, we used model fit values according to the following guidelines: As a criterion for the acceptance of the overall model, we assumed that χ^2/df should be ≤ 3.00 (Homburg and Giering, 1997). For the comparative fit index (CFI), values ≥ 0.90 indicate a good fit (Garson, 2009). Additionally, the root mean square error of approximation (RMSEA) should be ≤ 0.05 to be accepted as a good model fit or ≤ 0.08 for an acceptable model fit. The associated p -value of RMSEA must be ≥ 0.05 (Browne and Cudeck, 1993). All analyses were conducted using R-Studio.

Results

To what extent do German primary school teachers use learning data from digital learning platforms for individualization?

It was part of the survey to ask teachers about their previous and current usage of learning data from DLP for five different purposes of individualization (Figure 2). The self-reported usage is utilized as a dependent variable in the main analyses, but to answer research question 1 it is also examined descriptively.

Figure 2 shows that the teachers were divided into data users and non-users of roughly equal size across all five purposes. Up to 50% of the teachers had never used learning data from DLP for either of the specific purposes targeting individualization.



The majority of the teachers (58.5%) had used learning data from DLP to tailor instruction to individual students' needs. Of these, 48.9% claimed to currently use the learning data occasionally or regularly. Further, 46.7% of teachers reported that they currently used learning data from DLP to identify the learning needs of struggling students occasionally or regularly. Another 10.7% of participants had used the learning data at an earlier time. The third most common reason for using learning data from DLP was to set learning goals for individual students. Here, 11.0% of teachers had used learning data in the past and 43.0% currently did so. A total of 51.9% of teachers had previously used (10.3%) or currently use (41.6%) learning data from DLP to determine whether they needed to re-teach certain concepts and skills. Teachers were least likely to use learning data from DLP to identify the learning needs of students who were not struggling. Here, exactly 50.0% of the participants used learning data for this reason and correspondingly, the same number of participants did not. As with the previous purpose, 10.3% of teachers had used the learning data at an earlier time. The remaining teachers (39.7%) currently used the learning data at the time of the survey. The results showed that learning data from DLP were used by teachers to similar extents for different purposes.

What predicts teachers' intentions to use and the usage of learning data from digital learning platforms?

In preparation for the main analyses, we analyzed the descriptive statistics and reliability of all scales (Table 2). Since the reliability analyses yielded acceptable to excellent values for all scales, this issue is not considered further.

Results of the descriptive statistics showed that the mean score for *intention to use learning data from DLP* was $M = 2.75$. That is, on average, participants were slightly more likely to imagine using learning data from DLP within the school year

to design their lessons than to imagine not using them. A closer look at the evaluation of the items for the intention showed that 78% of teachers had an intention to use learning data from DLP. Thereby, the intention to use learning data differed among these teachers regarding its intensity. The remaining 22% of teachers had no intention to use learning data from DLP. An example item to assess the intention to use was "I plan to use learning data from digital learning platforms for my teaching within the current school term." Even though the items on the self-reported usage of learning data from DLP have already been considered in more detail (Section "To what extent do German primary school teachers use learning data from digital learning platforms for individualization?"), the mean value should also be mentioned here ($M = 2.07$). An example item for the usage scale was "I use data from digital learning platforms to set learning goals for individual students."

On average, the participants reported more positive than negative *attitudes toward learning data from DLP* ($M = 2.75$). This implies, for example, that teachers agreed that they find learning data from DLP useful. The mean for all participants regarding the scale *perceived behavioral control regarding learning data from DLP* was about moderate ($M = 2.53$). Thus, we could not make a clear determination about whether or not teachers might be able to deal with learning data from DLP. While some teachers stated that they are already able to use learning data from DLP other stated they cannot. For the *subjective norm regarding learning data from DLP* scale, the average answer showed a tendency toward negative response options ($M = 2.22$). This suggests that teachers were not particularly influenced by, for example, their colleagues to use learning data from DLP.

With regard to the *usage of digital media*, the teachers were asked about the use of ten different types of digital media in their lessons. We then calculated the sum score with a mean of $M = 5.19$. This showed that the teachers on average used about half of the digital media given. Teachers, on average, highly approved items regarding the *attitude toward digital media*

TABLE 2 Descriptive statistics and reliability analyses.

	Mean	SD	Range	Skew	Kurtosis	α
Usage of data from DLP	2.07	0.9	3	0.24	-1.22	0.9
Intention to use data from DLP	2.75	0.85	3	-0.35	-0.51	0.9
Attitude toward data from DLP	2.75	0.62	3	-0.3	0.21	0.9
Perceived behavioral control regarding data from DLP	2.53	0.75	3	-0.45	-0.44	0.9
Subjective norm regarding data from DLP	2.22	0.71	3	0.16	-0.29	0.8
Usage of digital media	5.19	1.83	10	0.02	-0.15	
Attitude toward digital media	3.03	0.51	2.88	-0.51	0.41	0.88
Practice	3.21	0.49	3	-1.41	3.77	0.75
Individual support	3.35	0.53	3	-0.82	0.69	0.75
Feedback	3.51	0.47	3	-1.04	1.86	0.77
Data-based instructional design	3.21	0.78	3	-1.25	1.19	0.94

$N = 272$.

($M = 3.03$). That is, teachers, on average, rather liked working with digital media in the classroom.

For most teachers, the didactical concepts considered, i.e., *practice*, *individual support*, *feedback*, and *data-based instructional design* certainly mattered in their teaching. Practicing was considered relevant by teachers on average ($M = 3.21$). For example, *practice* could be represented by the following item “I let my students practice with tasks where I can see particularly well whether the essentials have been understood.” Participants also perceived *individual support* to be relevant to their teaching with $M = 3.35$. For example, teachers occasionally to regularly gave their students tasks that fit their needs. Giving *feedback* appeared to be the most relevant didactic concept for teachers, as indicated by the mean of $M = 3.51$. On average, teachers occasionally to regularly told their students in which areas they could improve. For *data-based instructional design*, $M = 3.21$ showed the same

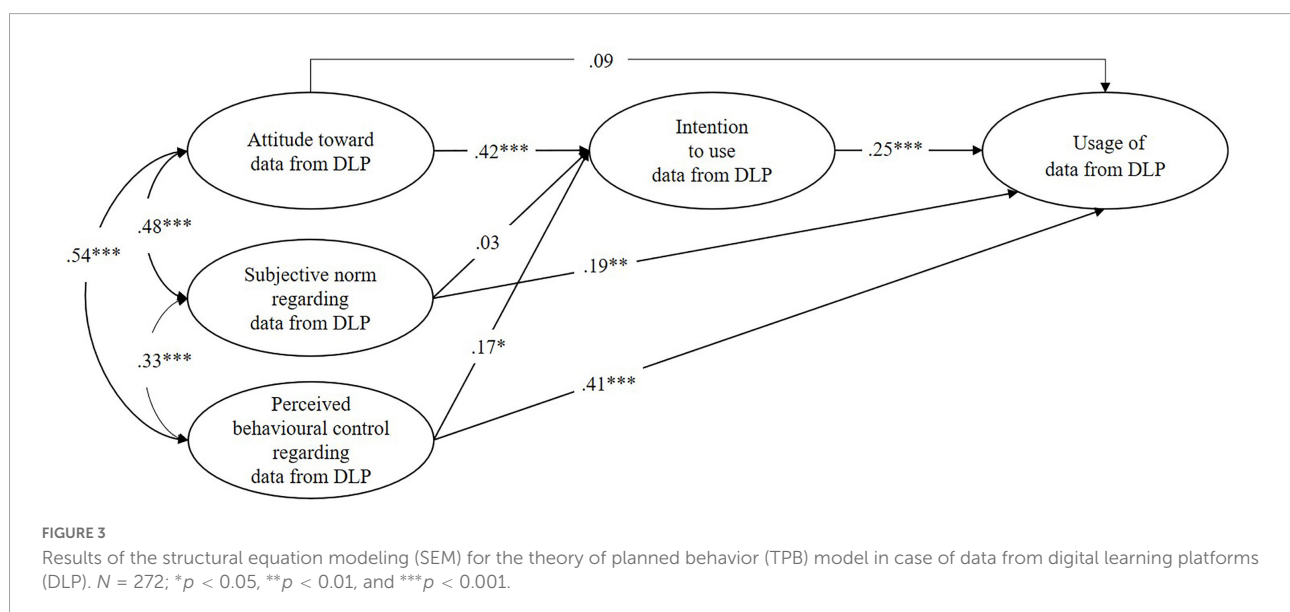
mean value as for practicing. Accordingly, teachers tended to use learning data occasionally; for example, as a basis for conversations with parents.

To finally answer the second research question, we predicted teachers’ intention to use learning data from DLP as well as their self-reported usage of learning data from DLP by firstly their *attitude*, *perceived behavioral control*, *subjective norm* and secondly as well by their *usage of digital media*, *attitude toward digital media*, *practice*, *individual support*, *feedback*, and *data-based instructional design*. The correlation matrix of all predictors with the intention to use and the usage of learning data from DLP is shown in Table 3. Cohen (1988) was followed in interpreting the correlation coefficients. Low to moderate significant correlations with intention to use learning data from DLP were found for all independent variables except for *practice*. *Attitude*, *perceived behavioral control*, *usage of digital media*, and *attitude toward digital media* showed moderate correlations

TABLE 3 Correlation matrix of all factors for all participants.

Variable	1	2	3	4	5	6	7	8	9	10	11
1 Usage of data from DLP	1										
2 Intention to use data from DLP	0.50***	1									
3 Attitude toward data from DLP	0.45***	0.49***	1								
4 Perceived behavioral control regarding data from DLP	0.54***	0.40***	0.48***	1							
5 Subjective norm regarding data from DLP	0.35***	0.23***	0.41***	0.28***	1						
6 Usage of digital media	0.37***	0.35***	0.21***	0.29***	0.10	1					
7 Attitude toward digital media	0.37***	0.39***	0.56***	0.40***	0.16**	0.31***	1				
8 Practice	0.02	0.07	0.10	−0.04	0.09	0.04	0.12*	1			
9 Individual support	0.17**	0.18**	−0.05	0.12*	0.07	0.18**	0.15**	0.26***	1		
10 Feedback	0.21***	0.22***	0.05	0.13*	0.08	0.10	0.10	0.20***	0.49***	1	
11 Data-based instructional design	0.18**	0.18**	0.13*	0.30***	−0.02	0.11	0.16**	0.33***	0.29***	0.33***	1

$N = 272$; * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.



with teachers' intentions to use learning data from DLP. A high positive correlation was found for intention to use learning data from DLP with usage of learning data from DLP. Additionally, the correlation matrix showed low to high correlations with usage of learning data from DLP and the other variables except for practice. Here, perceived behavioral control showed a high correlation with teachers' intentions to use learning data from DLP whereas moderate correlations were found for attitude, subjective norm, usage of digital media, and attitude toward digital media.

To examine the association of all variables in one model, we performed SEM. First, we considered the TPB model in its original form, but considered attitude and subjective norm as predictors for teachers' usage of data from DLP as well (Figure 3). Secondly, we extended the TPB model with additional variables: usage of digital media, attitude toward digital media, and didactical concepts (Figure 4). The influence of the additional variables was only tested regarding the intention to use data from DLP. Table 4 provides the standardized beta values of all relationships for both models.

Based on the confirmatory factor analysis, a good model fit could be established for Model 1: $\chi^2(242) = 385.65$, $\chi^2/df = 1.60$, $p \leq 0.001$. With a CFI = 0.95, the value represented a good model fit. The RMSEA = 0.05, with $p = 0.73$, and 90% CI [0.04, 0.06], could also be classified as good. Taking a closer look at the results of the first SEM, the intention to use data from DLP was mostly significantly predicted by teachers' attitudes toward learning data from DLP. Thus, teachers with a positive attitude toward the usage of learning data from DLP showed a higher intention to use it. Additionally, the perceived behavioral control regarding learning data from DLP also significantly predicted the intention to use learning data from DLP. Similarly, teachers who assessed their skills in using learning data from DLP as good showed a higher usage intention. In contrast, no significant associations were found for subjective norm regarding learning data from DLP. Regarding the usage of data from DLP, teachers' intentions to use learning data from DLP as well as their perceived behavioral control regarding learning data from DLP and subjective norm regarding learning data from DLP turned out as significant predictors. In this context, perceived behavioral control was most significant in explaining the model. Thus, we found that teachers use learning data from DLP when they perceive themselves as competent enough to do so or when other persons like colleagues influenced teachers' interest in using such learning data. With Model 1 we were able to explain 30% of the variance of the intention to use data from DLP and 51% of the variance of the usage of learning data from DLP.

Also for Model 2 a good model fit was established: $\chi^2(1277) = 1971.79$, $p \leq 0.001$, $\chi^2/df = 1.60$. Even though the CFI of 0.90 was a bit lower here, it could still be described as good. The RMSEA = 0.05, with $p = 0.99$, and 90% CI [0.04, 0.05], could also be classified as good. Model 2 explained 38% of the variance of the intention to use data from DLP and 52%

of the variance of the usage of learning data from DLP. As the additional predictors were tested only in relation to the intention to use learning data from DLP, only the consideration of ΔR^2 for intention was interesting. Following our default that an increase in explained variance becomes practically relevant only when $\Delta R^2 \geq 0.05$, a $\Delta R^2 = 0.08$ shows that the second model differed meaningfully from the first model. Therefore, Model 2 should be considered. In addition to the TPB variables, this model contained digital media (usage, attitude) and didactical concepts (practice, individual support, feedback, data-based instructional design). Again, attitude toward learning data from DLP most strongly predicted the intention to use learning data from DLP. In contrast to Model 1, perceived behavioral control regarding learning data from DLP did not predict teachers' intentions to use learning data from DLP. Of the added factors, usage of digital media was found to be a significant predictor of teachers' intentions to use learning data from DLP. Thus, in addition to teachers' attitudes, the usage of several types of digital media was predictive for their intention to use learning data from DLP. Other didactical concepts showed no significant association with the intention to use data from DLP. The previously identified predictors for the usage of data from DLP remained the same: intention to use data from DLP, perceived behavioral control regarding data from DLP, and subjective norm regarding data from DLP.

Discussion

Summary

The presented study provides a valuable insight into German primary school teachers' intention to use and usage of data from DLP. In this cross-sectional survey study, on the one hand, we were able to describe the usage of learning data from DLP for purposes of individualization. On the other hand, we predicted teachers' intention to use learning data from DLP as well as their usage with variables from the established TPB model (Model 1) as well as an extended TPB model (Model 2).

Regarding the first research question, about half of all participants indicated that they were already using learning data from DLP for various purposes of individualization. For example, identifying struggling students' learning needs led to great consent among the teachers. This emphasizes the added value of DBDM in the school context: Theoretical articles and empirical studies have cited the determination of appropriate instructional steps for students' individual learning needs as a reason for data usage (Mandinach and Gummer, 2016; Prenger and Schildkamp, 2018; Peters et al., 2021). Similarly, research on educational technologies has reported on the potential of DLP to provide information about students' needs from learning data (Greller et al., 2014; Schaumburg, 2021). Nevertheless, half of all teachers who participated in the survey did not use learning data

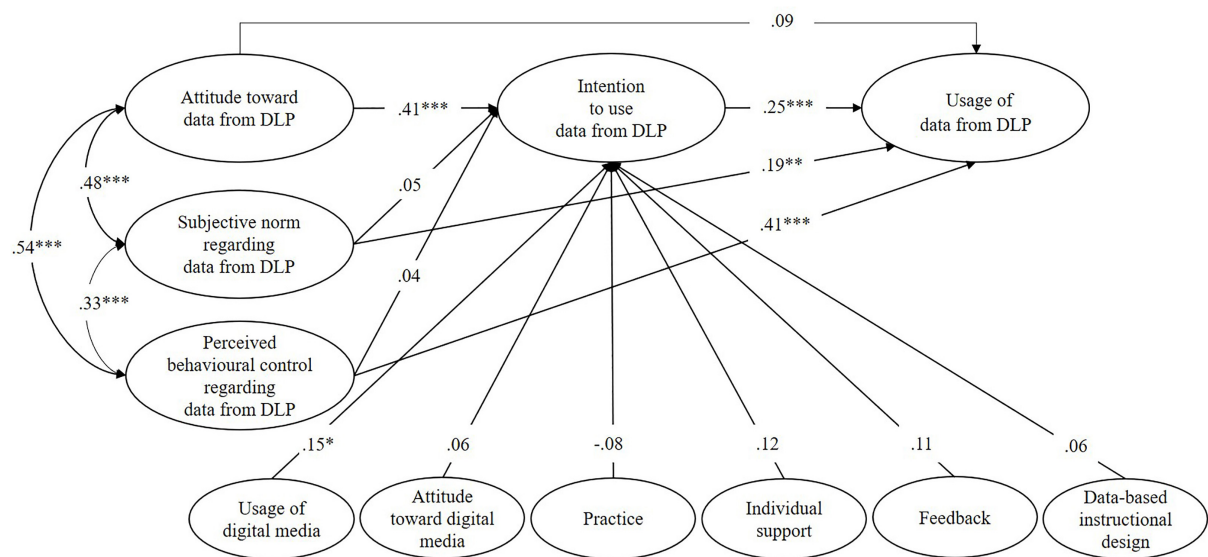


FIGURE 4

Results of the structural equation modeling (SEM) for the extended theory of planned behavior (TPB) model in case of data from digital learning platforms (DLP). $N = 272$; * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

TABLE 4 Standardized beta values of all relationships in the structural equation model.

	Model 1: TPB			Model 2: extended TPB		
	β	SE	R^2	β	SE	R^2
Intention to use data from DLP			0.30			0.38
Attitude toward data from DLP	0.42***	0.16		0.41***	0.19	
Perceived behavioral control regarding data from DLP	0.17*	0.10		0.04	0.11	
Subjective norm regarding data from DLP	0.03	0.13		0.05	0.13	
Usage of digital media				0.15*	0.03	
Attitude toward digital media				0.06	0.11	
Practice				-0.08	0.17	
Individual support				0.12	0.18	
Feedback				0.11	0.22	
Data-based instructional design				0.06	0.09	
Usage of data from DLP			0.51			0.52
Intention to use data from DLP	0.25***	0.07		0.25***	0.07	
Attitude toward data from DLP	0.09	0.16		0.09	0.15	
Perceived behavioral control regarding data from DLP	0.41***	0.10		0.41***	0.09	
Subjective norm regarding data from DLP	0.19**	0.13		0.19**	0.12	

$N = 272$; * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

from DLP. A non-use of learning data was also found in other studies (Kippers et al., 2018; Blumenthal et al., 2021). For this reason, it was important to further investigate the reasons for the intention to use and usage of learning data from DLP.

With regard to the second research question on predictors of intention to use and usage of learning data from DLP, teachers' attitudes toward learning data from DLP proved to be the most relevant predictor for intention to use learning

data from DLP in both models. Therefore, teachers need a positive mindset about learning data from DLP in order to consider using them. Teo and Tan (2012) also found the highest influence of attitude as a factor of the TPB model when they predicted teachers' intentions to use technology in school. Likewise, Blumenthal et al. (2021) identified attitude toward data—independently of TPB—as an important predictor for the intention to use data for educational decisions. In contrast to

previous studies, however, subjective norm had no effect on the intention to use learning data from DLP (Teo and Tan, 2012; Hellmich et al., 2019; Knauder and Koschmieder, 2019). This might be explained by the fact that teachers in Germany are quite independent in their lesson planning and often do not receive regulations regarding the choice of their methods (Kerres, 2020). The irrelevance of subjective norm changed when considering the TPB variables in terms of the usage of learning data from DLP. Here, subjective norm regarding learning data from DLP significantly predicted teachers' usage of learning data from DLP. In return, the attitude toward learning data from DLP had no influence on the explanation of the usage of learning data from DLP. The relevance of perceived behavioral control regarding learning data from DLP and the irrelevance of attitude toward learning data from DLP to the usage of learning data from DLP is consistent with the findings of Knauder and Koschmieder (2019) on the consideration of TPB with respect to individualized instructional design but is also in contrast to the findings of other studies (Prenger and Schildkamp, 2018; Hellmich et al., 2019). As expected, intention to use learning data from DLP had a significant influence on the usage of learning data from DLP. Nevertheless, an intention-behavior gap is evident here as well (Ajzen, 1991; Sheeran, 2002): More teachers have the intention to use, but fewer actually realize the usage of learning data from DLP. This may be due to the fact that it takes more than just a positive attitude to use it. It also requires competencies—expressed here in perceived behavioral control—that must first be acquired. The addition of further variables led to a meaningful improvement of the model, but only the previous usage of digital media could be identified as a significant predictor of the intention to use learning data from DLP. Therefore, it is helpful for teachers to be able to imagine the usage of learning data from DLP if they have already gained experience with other digital media. From this we can assume that there would also be a significant association between the intention to use or the usage of the DLP and the intention to use learning data from DLP. The extension of a predictive model to include didactical concepts, like the importance of feedback or the usage of data-based instructional design, as proposed by Tappe (2017) and Gellerstedt et al. (2018), yielded no success in this study. We conclude that even though the TPB model proved to be very robust and the influence of additional predictors was small, it seems useful to consider the teacher's instructional context when explaining teachers' intentions to use learning data from DLP in future studies.

Strengths and limitations

To the best of our knowledge, this is the first study to examine primary teachers' intention to use and usage of learning data from DLP in the context of individualization in

Germany. In this context, an already established theoretical model proved useful in the cross-sectional survey study and was tested with additional factors. Nevertheless, there are some limitations to this study.

In this study, the TPB model was considered in terms of both teachers' intentions and usage of learning data from DLP. However, it is worth noting that teacher respondents were only surveyed at a single point in time, as it is desirable to observe the intentional and behavioral change over a certain time between the first and the second measurement. Accordingly, the results should be confirmed in a longitudinal survey. Nevertheless, for comparability with other studies of the TPB, both intention and usage were included in our analyses. Moreover, the usage—as well as the other items—was only self-reported by the teachers, thus there is a possibility of distortion. The real usage of learning data from DLP, how it is designed, and if it is beneficial for learning of students is left unanswered and should be subject of further research.

Further, although all primary school teachers in Bremen, Hamburg, Mecklenburg-Western Pomerania, Lower Saxony, and Schleswig Holstein were contacted via their schools and invitations were issued on social media to participate in the survey, only a small number of primary school teachers took part in the survey study. Nevertheless, this number met the previously calculated sample size and analyses were conducted. In addition, it can be assumed that the sample is characterized by media-literate teachers, as the respondents were recruited via e-mail and social media and the questionnaire was conducted online. The frequency of usage might be overestimated.

Outlook

This study was able to explain primary school teachers' intention to use and the usage of learning data from DLP especially for individualization. Doing so, this study contributes to the growing body of research on the potentials of DBDM and Learning Analytics in the context of inclusive schooling. Nevertheless, further empirical research is needed based on these findings. We have already been able to explain part of the intention to use and the usage of data from DLP, however, some reasons for the (non) use still remain unexplained. These need to be investigated in further studies. In this context, we also recommend qualitative studies, for example interviews with primary school teachers, in order to elaborate further relevant factors. Since students in primary schools are particularly heterogeneous, we focused on primary school teachers. However, an investigation of the model would also be interesting for secondary school teachers. In addition to our findings regarding predictors of intention to use and usage, it would be valuable to better understand what motivates teachers to use or not use learning data from DLP. Moreover, it would

be interesting to find out more about how teachers use learning data from DLP, especially in the context of individualization, and whether this use has an impact on learning effectiveness. To this end, qualitative studies like interviews, school-based observation studies, or additional quantitative studies are desirable. Furthermore, it would certainly be worthwhile to take a closer look at teachers' competencies for using learning data from DLP and to investigate the influence on their intention to use as well as their usage.

The results of this study indicate that it is also necessary to consider its implications for teacher education and training. Consideration needs to be given to how teachers' attitudes toward the usage of learning data from DLP, as well as their perceived behavioral control, can be fostered in teacher trainings to increase their usage of learning data from DLP. Because DBDM, especially in DLP contexts, can support teachers in establishing individualized learning opportunities, this can help to meet the needs of all students best.

Data availability statement

The raw data supporting the conclusions of this article will be provided by the authors upon request.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of the Leuphana University Lüneburg. The participants provided their written informed consent to participate in this study.

Author contributions

AH had the main role in writing the original draft. LK, PK, and DL contributed in editing the manuscript. AH and

LK mainly conducted the analyses. All authors contributed to funding acquisition and project administration and approved the submitted version.

Funding

This study was funded by the Open Access Publication Fund of the Leuphana University Lüneburg. This study developed within the CODIP project. CODIP is funded by the Quality Initiative Teacher Training (Qualitätsoffensive Lehrerbildung), a joint initiative of Federal Government and the German states. The financial means were provided by the Federal Ministry of Education and Research (BMBF) (Support code: 01JA2002).

Acknowledgments

We thank the teachers from our development team for their help in revising the questionnaire and our student assistants for their support in data evaluation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decision Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T
- Anderson, S., Jungjohann, J., and Gebhardt, M. (2020). Effects of using curriculum-based measurement (CBM) for progress monitoring in reading and an additive reading instruction in second classes. *Zeitschrift für Grundschulforschung* 13, 151–166. doi: 10.1007/s42278-019-00072-5
- Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., et al. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente [Teachers' professional knowledge, cognitively activating mathematics instruction, and the development of mathematical competence (COACTIV): documentation of survey instruments]*. Berlin: Max-Planck-Institut für Bildungsforschung.
- Blumenthal, S., Blumenthal, Y., Lembke, E. S., Powell, S. R., Schultze-Petzold, P., and Thomas, E. R. (2021). Educator Perspectives on Data-Based Decision Making in Germany and the United States. *J. Learn. Disabilities* 54, 284–299. doi: 10.1177/0022219420986120
- Böhme, R., Munser-Kiefer, M., and Prestidge, S. (2020). Lernunterstützung mit digitalen Medien in der Grundschule: Theorie und Empirie zur Wirkweise zentraler Funktionen und Gestaltungsmerkmale [Support Learning with Digital Media in Elementary School: Theory and Empirical Evidence on the Effectiveness of Key Functions and Design Features]. *Zeitschrift für Grundschulforschung* 13, 1–14. doi: 10.1007/s42278-019-00066-3
- Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit," in *Testing Structural Equation Models*, eds K. A. Bollen and J. S. Long (Newbury Park, CA: Sage Publications), 136–162.

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: L. Erlbaum Associates.
- Daniela, L., and Rüdolf, A. (2019). "Learning platforms: how to make the right choice," in *Didactics of Smart Pedagogy: Smart Pedagogy for Technology Enhanced Learning*, ed. L. Daniela (Cham: Springer), 191–209. doi: 10.1007/978-3-030-01551-0
- Eid, M., Gollwitzer, M., and Schmitt, M. (2017). *Statistik und Forschungsmethoden [Statistics and research methods]*. Weinheim, Basel: Beltz.
- Faustmann, G., Lemke, C., Kirchner, K., and Monett, D. (2019). "Which factors make digital learning platforms successful," in *Proceedings of the 13th Annual International Technology, Education and Development Conference*, Velancia, 6777–6786. doi: 10.21125/inted.2019.1651
- Federal Statistical Office of Germany (2022). *Allgemeinbildende Schulen: Fachserie 11 Reihe 1 – Schuljahr 2020/2021 [General education schools: Subject-matter series 11 series 1 – School year 2020/2021]*. Available online at: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Schulen/Publikationen/_publikationen-innen-schulen-allgemeinbildende.html (Accessed on March 28, 2022).
- Fishbein, M., and Ajzen, I. (1975). *Belief, attitude, Intention and Behaviour: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.
- FitzGerald, E., Jones, A., Kucirkova, N., and Scanlon, E. (2018). A literature synthesis of personalised technology-enhanced learning: what works and why. *Res. Learn. Technol.* 26, 1–16. doi: 10.25304/rlt.v26.2095
- Garson, G. D. (2009). *Structural Equation Modeling*. Asheboro, NC: Statistical Associates Publishers.
- Gellerstedt, M., Babaheidari, S. M., and Svensson, L. (2018). A first step towards a model for teachers' adoption of ICT pedagogy in schools. *Heliyon* 4, 1–17. doi: 10.1016/j.heliyon.2018.e00786
- Greller, W., and Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educ. Tech. Soc.* 15, 42–57.
- Greller, W., Ebner, M., and Schön, M. (2014). Learning analytics: From theory to practice – data support for learning and teaching. *Communicat. Comp. Inform. Sci.* 439, 79–87. doi: 10.1007/978-3-319-08657-6_8
- Hellmich, F., Löper, M. F., and Görel, G. (2019). The role of primary school teachers' attitudes and self-efficacy beliefs for everyday practices in inclusive classrooms: A study on the verification of the "Theory of Planned Behaviour". *J. Res. Special Educ. Needs* 19, 36–48. doi: 10.1111/1471-3802.12476
- Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., and Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Comput. Educ.* 153, 1–25. doi: 10.1016/j.compedu.2020.103897
- Holmes, W., Anastopoulou, S., Schaumburg, H., and Mavrikis, M. (2018). *Technology-enhanced Personalised Learning. Untangling the Evidence*. Stuttgart: Robert Bosch Stiftung.
- Homburg, C., and Giering, A. (1997). Konzeptualisierung und Operationalisierung komplexer Konstrukte: Ein Leitfaden für die Marketingforschung [Conceptualization and operationalization of complex constructs – A guideline for marketing research]. *Marketing* 18, 5–24.
- Ifenthaler, D., and Drachsler, H. (2020). "Learning Analytics: Spezielle Forschungsmethoden in der Bildungstechnologie [Learning analytics: special research methods in educational technology]," in *Handbuch Bildungstechnologie [Educational Technology Handbook]*, eds H. Niegemann and A. Weinberger (Berlin: Springer), 515–534. doi: 10.1007/978-3-662-54368-9_42
- Institute for Quality Development Hessen (2012). *Hessischer Referenzrahmen Schulqualität: Dokumentation der Fragebogen [Hessian Reference Framework for School Quality: Documentation of the questionnaires]*. Wiesbaden: Institute for Quality Development (IQ) Hessen.
- Jäger, R. S., and Helmke, A. (2008). *Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext (MARKUS) (Version 1) [Mathematics Survey Rhineland-Palatinate: Competencies, Instructional Characteristics, School Context (MARKUS) (Version 1)]*. Berlin: IQB – Institute for Educational Quality Improvement, https://doi.org/10.5159/IQB_MARKUS_v1
- Kerres, M. (2018). *Mediendidaktik. Konzeption und Entwicklung digitaler Lernangebote [Media didactics. Conception and development of digital learning offers]*. Berlin: De Gruyter, doi: 10.1515/9783110456837
- Kerres, M. (2020). Against All Odds: Education in Germany Coping with Covid-19. *Postdigital Sci. Educ.* 2, 690–694. doi: 10.1007/s42438-020-00130-7
- Keuning, T., van Geel, M., Visscher, A., and Fox, J.-P. (2019). Assessing and validating effects of a data-based decision-making intervention on student growth for mathematics and spelling. *J. Educ. Measure.* 56, 757–792. doi: 10.1111/jedm.12236
- Kippers, W. B., Wolterinck, C. H. D., Schildkamp, K., and Poortman, C. L. (2018). Teachers' views on the use of assessment for learning and data-based decision making in classroom practice. *Teach. Teach. Educ.* 75, 199–213. doi: 10.1016/j.tate.2018.06.015
- Knauder, H., and Koschmieder, C. (2019). Individualized student support in primary school teaching: A review of influencing factors using the Theory of Planned Behavior (TPB). *Teach. Teach. Educ.* 77, 66–76. doi: 10.1016/j.tate.2018.09.012
- Knickenberg, M., Zurbriggen, C. L. A., Venetz, M., Schwab, S., and Gebhardt, M. (2020). Assessing dimensions of inclusion from students' perspective: measurement invariance across students with learning disabilities in different educational settings. *Eur. J. Special Needs Educ.* 35, 287–302. doi: 10.1080/08856257.2019.1646958
- Krein, U., and Schiefner-Rohs, M. (2021). Data in Schools: (Changing) Practices and Blind Spots at a Glance. *Front. Educ.* 6:672666. doi: 10.3389/feduc.2021.672666
- Mandinach, E. B., and Gummer, E. S. (2013). A Systemic View of Implementing Data Literacy in Educator Preparation. *Educ. Res.* 42, 30–37. doi: 10.3102/0013189X12459803
- Mandinach, E. B., and Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teach. Teach. Educ.* 60, 366–376. doi: 10.1016/j.tate.2016.07.011
- Mandinach, E. B., and Schildkamp, K. (2020). Misconceptions about data-based decision making in education: An exploration of the literature. *Stud. Educ. Eval.* 2020:100843. doi: 10.1016/j.stueduc.2020.100842
- Molenaar, I., and Knoop-van Campen, C. A. N. (2017). *Teacher Dashboards in Practice: Usage and Impact, in Data Driven Approaches in Digital Education: EC-T&L 2017. LNCS 10474*. Basel: Springer Cham, 15–138. doi: 10.1007/978-3-319-66610-5_10
- Molenaar, I., and Knoop-van Campen, C. A. N. (2018). How teachers make dashboard information actionable. *IEEE Transac. Learn. Technol.* 12, 347–355. doi: 10.1109/TLT.2018.2851585
- Moore, R., and Shaw, T. (2017). *Teachers' Use of Data: An Executive Summary*. Available online at: <https://www.act.org/content/dam/act/unsecured/documents/R1661-teachers-use-of-data-2017-12.pdf> (accessed July 29, 2021).
- Nattland, A., and Kerres, M. (2009). "Computerbasierte Medien im Unterricht [Computer-based media in the classroom]," in *Handbuch Unterricht [Teaching Handbook]*, eds K.-H. Arnold, J. Wiechmann, and U. Sandfuchs (Bad Heilbrunn: Klinkhardt), 317–323.
- Nistor, N. (2020). "Akzeptanz von Bildungstechnologien [Acceptance of Education Technology]," in *Handbuch Bildungstechnologie [Educational Technology Handbook]*, eds H. Niegemann and A. Weinberger (Berlin: Springer), 535–545. doi: 10.1007/978-3-662-54368-9_46
- Peters, M. T., Förster, N., Hebbeker, K., Forthmann, B., and Souvignier, E. (2021). Effects of data-based decision-making on low-performing readers in general education classrooms: cumulative evidence from six intervention studies. *J. Learn. Disabilities* 54, 334–348. doi: 10.1177/00222194211011580
- Petko, D. (2014). *Einführung in die Mediendidaktik: Lehren und Lernen mit digitalen Medien [Introduction to Media Didactics: Teaching and learning with digital media]*. Weinheim: Beltz.
- Petko, D., Prasse, D., and Cantieni, A. (2018). The interplay of school readiness and teacher readiness for educational technology integration: A structural equation model. *Comput. Sch.* 35, 1–18. doi: 10.1080/07380569.2018.1428007
- Pierce, R., Chick, H., and Gordon, I. (2013). Teachers' perceptions of the factors influencing their engagement with statistical reports on student achievement data. *Aust. J. Educ.* 57, 237–255. doi: 10.1177/0004944113496176
- PISA (2017). *Teacher Questionnaire for Pisa 2018: General Teacher. Main Survey Version*. Available online at: <https://www.oecd.org/pisa/data/2018database> (Accessed on July 29, 2021).
- Prenger, R., and Schildkamp, K. (2018). Data-based decision making for teacher and student learning: a psychological perspective on the role of the teacher. *Educ. Psychol.* 38, 734–752. doi: 10.1080/01443410.2018.1426834
- Reinhold, F., Hoch, S., Werner, B., Richter-Geibert, J., and Reiss, K. (2020). Learning fractions with and without educational technology: What matters for high-achieving and low-achieving students? *Learn. Instr.* 65, 1–19. doi: 10.1016/j.learninstruc.2019.101264
- Schaumburg, H. (2021). Personalisiertes Lernen mit digitalen Medien als Herausforderung für die Schulentwicklung: Ein systematischer Forschungsüberblick [Personalized Learning with Digital Media as a Challenge for School Development: A Systematic Research Review]. *Medienpädagogik* 41, 134–166. doi: 10.21240/mpaed/41/2021.02.24.X

- Schaumburg, H., and Prasse, D. (2019). *Medien und Schule: Theorie – Forschung – Praxis [Media and school: Theory – Research – Practice]*. Bad Heilbrunn: Julius Klinkhardt.
- Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educ. Res.* 61, 257–273. doi: 10.1080/00131881.2019.1625716
- Schildkamp, K., and Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teach. Teach. Educ.* 26, 482–496. doi: 10.1016/j.tate.2009.06.007
- Schmid, U., Goertz, L., and Behrens, J. (2017). *Monitor Digitale Bildung: Die Schulen im digitalen Zeitalter [Digital Education Monitor: Schools in the digital age]*. Gütersloh: Bertelsmann Stiftung, doi: 10.11586/2017041
- Schumacker, R. E., and Lomax, R. G. (2010). *A Beginner's Guide to Structural Equation Modeling*. New York, NY: Routledge.
- Schwab, S., Hellmich, F., and Görel, G. (2017). Self-efficacy of prospective Austrian and German primary school teachers regarding the implementation of inclusive education. *J. Res. Special Educ. Needs* 17, 205–217. doi: 10.1111/1471-3802.12379
- Sheeran, P. (2002). Intention—behavior relations: A conceptual and empirical review. *Eur. Rev. Social Psychol.* 12, 1–36. doi: 10.1080/14792772143000003
- Souvignier, E., Förster, N., Hebbeker, K., and Schütze, B. (2021). “Using Digital Data to Support Teaching Practice – quop: An Effective Web-Based Approach to Monitor Student Learning Progress in Reading and Mathematics in Entire Classrooms,” in *International Perspectives on School Settings, Education Policy and Digital Strategies: A Transatlantic Discourse in Education Research*, eds A. Wilmers and S. Jörnitz (Leverkusen: Verlag Barbara Budrich), 283–298. doi: 10.2307/j.ctv1gbrzf4.20
- Tappe, E. -H. (2017). *Lernen durch Mediengestaltung: Entwicklung eines Konzeptes zur Unterstützung mediendidaktischer Lehre im Schulalltag [Learning through Media Design: Development of a Concept to Support Media Didactic Teaching in Everyday School Life]*. [dissertation]. Münster: Westfälische Wilhelms-Universität.
- Teo, T., and Tan, L. (2012). The Theory of Planned Behavior (TPB) and Pre-Service Teachers' technology acceptance: A validation study using structural equation modeling. *J. Technol. Teach. Educ.* 20, 89–104.
- Tondeur, J., Aesaert, K., Prestridge, S., and Consuegra, E. (2018). A multilevel analysis of what matters in the training of pre-service teacher's ICT competencies. *Comput. Educ.* 122, 32–42. doi: 10.1016/j.compedu.2018.03.002
- Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. New York, NY: CRC Press Taylor & Francis Group.
- Vanbecelaere, S., Cornillie, F., Depaepe, F., Guerrero, R. G., Mavrikis, M., Vasalou, M., et al. (2020). “Technology-mediated personalised learning for younger learners,” in *Proceedings of the 2020 ACM Interaction Design and Children Conference: Extended Abstracts*, eds S. Price, E. Rubegni, and A. Vasalou (New York, NY: ACM), 126–134. doi: 10.1145/3397617.3398059
- Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Q.* 27, 425–478.
- Wayman, J. C., Wilkerson, S. B., Cho, V., Mandinach, E. B., and Sopovitz, J. A. (2016). *Guide to using the Teacher Data Use Survey (REL 2017–166)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia.
- Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., and Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assess. Educ.* 28, 228–260. doi: 10.1080/0969594X.2021.1884



OPEN ACCESS

EDITED BY

Farah El Zein,
Emirates College for Advanced
Education, United Arab Emirates

REVIEWED BY

Ellen B. Mandinach,
WestEd, United States
S. E. Mol,
Leiden University, Netherlands

*CORRESPONDENCE

David Scheer
david.scheer@ph-ludwigsburg.de

SPECIALTY SECTION

This article was submitted to
Special Educational Needs,
a section of the journal
Frontiers in Education

RECEIVED 13 April 2022

ACCEPTED 09 August 2022

PUBLISHED 07 September 2022

CITATION

Jungjohann J, Gebhardt M and
Scheer D (2022) Understanding
and improving teachers' graph literacy
for data-based decision-making via
video intervention.
Front. Educ. 7:919152.
doi: 10.3389/feduc.2022.919152

COPYRIGHT

© 2022 Jungjohann, Gebhardt and
Scheer. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Understanding and improving teachers' graph literacy for data-based decision-making via video intervention

Jana Jungjohann¹, Markus Gebhardt¹ and David Scheer^{2*}

¹Department of Education Science II, University of Regensburg, Regensburg, Germany, ²Faculty for Special Needs Education, Ludwigsburg University of Education, Ludwigsburg, Germany

In the educational context, graph literacy describes the competence to read, comprehend, and interpret formative assessment data in terms of data-based decision-making (DBDM) in order to derive and justify individual adaptations of instruction based on them. Since misconceptions may arise in predicting a future learning progress due to the characteristics of the data base as well as the approach to graph literacy, effective supports are needed, especially for inexperienced teachers. We present two interrelated studies to broaden the field of support in graph literacy. In Study I, graph literacy procedures are collected from $N = 196$ university student teachers using an online survey that includes six case vignettes with learning progress prediction tasks. Results show that both regular and special education student teachers intuitively neglect important data points in interpretation and they do not use a consistent strategy in prediction across the case vignettes (Fleiss' $\kappa = 0.071$; $p < 0.001$). Building on the results of Study I, a 3-min video intervention for linear trend identification using Tukey Tri-Split was developed. Study II tested the efficacy of the video intervention on the accuracy of future learning progress among student teachers and in-service teachers ($N = 198$) using randomized group assignment compared to a brief text hint. With a large effect size of Cohens' $f = 0.39$, the video instruction shows high efficacy compared to the text hint. The increasing importance of DBDM in inclusive and special education is discussed.

KEYWORDS

data-based decision-making (DBDM), formative assessment, graph literacy, instructional effectiveness, progress monitoring, teacher education, video-based intervention

Introduction

Teachers' graph literacy is a widely neglected skill that influences decision-making performance (Okan et al., 2012; Oslund et al., 2021). It matters as a core component of data literacy for all teachers. Following Mandinach and Gummer (2016), it is defined as the need of knowledge of "how to use data displays because data are often graphically depicted, in chart, tables, graphs, and other displays" (p. 371). This definition is broad and refers to both qualitative and quantitative data generated in the school context. In inclusive and special education, the use of formative assessment is widespread in order to use quantitative data to discover learning problems and to adapt instruction to meet children's needs in the sense of data-based decision-making (DBDM; Espin et al., 2021). Therefore, in this paper, graph literacy is considered in terms of quantitative data only.

Formative assessments are used primarily in multi-tiered systems of support in different learning areas such as reading, writing and mathematics (Fien et al., 2021) and as a supplement to cross-sectional status tests in the area of instruction planning around the world (e.g., Fuchs, 2017; Jungjohann et al., 2018a; Ahmed, 2019). Especially in school systems without an implemented multi-tiered system of supports such as Germany, there is a lack of standardized and effective training and further education for teachers (Blumenthal et al., 2021).

The goal of using formative assessment data is for teachers to make informed decisions based on student data to achieve a better fit between learning needs and instruction and therefore to achieve a higher students' achievement outcome. For this, teachers collect ongoing diagnostic data by using formative tests to measure learning growth and identify students who need support at Tier 2 or Tier 3 (Lane et al., 2014). In the formative assessment approach, formal and informal formative measures can be distinguished. Formal tests produce mostly quantitative data from standardized assessments and informal tests collect both qualitative and quantitative data from homework assignments or in-class activities. Standardized tests for learning progress monitoring are used at high frequency up to weekly during lessons, take only a few minutes, and are based on specific quality criteria (Good and Jefferson, 1998). The tests must be reliable, on the one hand, and short enough, on the other hand, to use little learning time, be easy to use in the classroom, and not overload the students (Schurig et al., 2021). In most cases, these quantitative measures are designed and scored as simple speed tests. This means that the students work on as many tasks as they can manage in the fixed test time (Kubinger, 2005). The outcome variable is traditionally the sum of all correctly solved tasks. It is usually visualized in a computer-based or drawn by hand graph as the student's learning growth (i.e., slope or rate of improvement) with the assumption that visual representations of numeric data facilitate inferences about conceptual relationships (Kosslyn,

2006). Therefore, on the graph's x-axis, the progression over time as the number of school weeks is shown. Here, teachers can read the single measurement points and the time intervals of the learning progress tests performed (Jungjohann et al., 2018c). The y-axis shows the outcome variable. If several test results are available, they are connected with a line to form a learning slope. The slope is one key component of the output of progress monitoring tests because it alerts teachers when students are not progressing successfully (Fuchs and Fuchs, 2001; Stecker et al., 2008). To prevent potential school failure, teachers use the measured outcome for both justifying adaptations to individual instruction and predicting the most likely future learning growth slope.

The use of formative assessment is particularly effective in supporting at-risk students and with difficulties in learning such as students with special educational needs (Bennett, 2011) because students achieve higher when their learning growths are monitored and reported to the teacher (Carlson et al., 2011; Anderson et al., 2020; McMaster et al., 2020). However, DBDM is only sporadically used by teachers and has not yet been adequately supported, required, and encouraged in many school systems (Blumenthal et al., 2021). Despite the positive impact of DBDM on student learning, Gleason et al. (2019) demonstrated that it takes a lot of effort to motivate in-service teachers to use DBDM. In their intervention study, 470 teachers from 102 American schools from 12 districts participated. Although they initiated an extensive support for DBDM (i.e., hiring data coaches, informing teachers in data-driven instruction, initiating data-focused teacher team meetings) on school level, no increase of teachers' data use or a change in teachers' instructional practices could be observed. A complementary research approach at teacher level focuses on promoting accurate visual analysis to strengthen the impact of DBDM. On the one hand, researchers try to better understand teachers' understanding of progress monitoring graphs (Espin et al., 2017; Klapproth, 2018) and, on the other hand, support measures for improved interpretation and prediction of learning are developed (Wagner et al., 2017; van den Bosch et al., 2019). It is necessary to take a closer and simultaneous look at both teachers' approach to interpreting the data in the graphs and the design of supporting materials. This is because only with a firm understanding of the current approach could support measures for teachers be developed and used effectively.

Graph literacy

For graph literacy, also known as graph comprehension, no universal definition exists. In accordance with Oslund et al. (2021) and with regard to quantitative progress monitoring data, graph literacy can be understood as multiple levels of reading and comprehension data and interpreting the graphs' slope. For evaluating the effectiveness of instructional programs,

teachers do multiple steps. They interpret the actual learning development of individuals based on the progress monitoring data, link the individual growth with the instructional programs, and predict a possible growth. Zeuch et al. (2017) describe three levels of graph literacy: (1) *reading the data*: notice the relevant data points and trends, (2) *reading between the data*: recognize relations between the developments of sub-competencies, and (3) *read beyond the data*: infer assumptions about further progress, possible deficits, and adequate instructional strategies for students. These three levels are hierarchical and build on each other. Reading the data level is of particular importance in graph literacy, as it is the foundation for interpretation. In this level, teachers decide which parts of the available data base they will include in their interpretation and which strategy they will use to arrive at their prediction. To reach the highest level of graph literacy to take full advantage of the potential of learning progress data, teachers must still combine all individual levels.

Graph literacy in the sense of DBDM is complex and requires teachers' diagnostic and pedagogical competencies to provide overlooking individual learning difficulties and profiles. There is a large evidence that teachers have multiple difficulty using quantitative data to inform and guide their instruction, especially in the areas of reading data concerning the data base under consideration (e.g., Keuning et al., 2017; Gesel et al., 2021). Teachers can have difficulties on the lowest interpretation level, when they focus on a single or irrelevant data points and disregard important information. Additionally, visual support within the graphs (e.g., linear trend line, goal lines, vertical border lines between interventions) can even distract the interpretation (Newell and Christ, 2017). On the intermediate interpretation level, data characteristics bias data prediction (Klapproth, 2018). For example, extreme values, high data variability, and a flat improvement cause a more positive prediction.

In addition to graph's layout, data base under consideration and interpretation strategies, the viewer's prior knowledge and the educational content of the graph can challenge the interpretation (Glazer, 2011). For instance, Wagner et al. (2017) compared the graph interpretation strategies of student teachers in special education and scholars in DBDM with think aloud procedure twice, just before and after completing student teaching. Measured by the number of words and statements, student teachers interpreted the graphs with lower coherence, specificity, reflection and accuracy than experts. The results suggest that graph literacy can be increased by specific training. In addition, Oslund et al. (2021) examined the influence that affective variables (i.e., teacher experience, hours of teacher training in data use and response-to-intervention approaches, and confidence on graph literacy) have on DBDM in the context of reading fluency tests. With a sample of 309 K-12 teachers, they found that both teachers' experience and confidence had an effect on teachers' graph literacy while the variable hours of teacher training did not. These results strengthen the

assumption that graph interpretation can be trained on the basis of teachers' prior knowledge and experience, and that training success depends on content rather than time.

Intervention on graph reading

To ensure a competent use of progress monitoring graphs, teachers need effective support (Ardoin et al., 2013). Gesel et al. (2021) concluded in their meta-analyses on the impact of DBDM training (i.e., data collection, analysis, data-based adaptations) targeting on teacher-level DBDM outcome (i.e., DBDM knowledge, skill and/or self-efficacy) a mean effect size of $g = 0.57$ for student teachers and in-service K-12 teachers in different school settings. Compared to Gleason et al. (2019) findings that even extensive support for DBDM including multiple aspects of teachers' trainings related to DBDM does not lead to changes in teacher behavior, the effects Gesel et al. (2021) found seem promising. These findings suggest that individual interventions can increase teachers' understanding of learning progress data. However, Espin et al. (2021) noted that teacher training and supporting materials must explicitly focus on DBDM procedures for positive effects.

In the context of teacher professional development, science video-based interventions are often used for multiple reasons. Video-based interventions are effective, have a simple and flexible handling and can have low production costs. Boy et al. (2020) distinguish four types of science videos: presentation videos, expert videos, animation videos and narrative explanatory videos. They investigated differences in knowledge transfer by multiple-choice tests and revealed a small benefit of narrative explanatory and animation videos. Animation videos present the relevant information in an audio channel through an off-screen invisible narrator and punctuate the information with artificial moving images. The advantage of these videos is that they can be very short and are suitable for explaining simple facts. Narrative explanatory videos are much more complex. They combine moving images with moderation or interview elements to provide comprehensive answers to complex questions. van den Bosch et al. (2019) used animated videos as video interventions in which a teacher presents the case of Sander and his reading difficulties. Their study indicated that teachers' graph literacy can be improved by animated video interventions. They used a pre-post-design with three different animated video interventions focused on basic knowledge, interpretation knowledge, and interpretation and linking knowledge that lasted between 20 and 45 min to deliver multiple instructional approaches and one control condition. Graph literacy was measured by a graph description task. In this task, teachers were asked to say out loud everything they saw in the graphs and interpret them as if they were talking to parents. With all three animated video intervention conditions teachers improved their graph literacy.

However, little is known, especially in the German-speaking education system, about the strategies teachers use to approach the interpretation of learning progression graphs or about their prior knowledge in this regard (Blumenthal et al., 2021). At the same time, we argue that an essential basic skill for all three levels of graph interpretation is the recognition and continuation of linear development trends. Despite this, it is not practical for teachers to estimate slope coefficients based on linear regressions in the context of data-based decisions in everyday pedagogy, especially as teacher training usually does not contain sufficient statistics courses for using robust regression. Rather, what is needed is a graphically implementable method that can be quickly learned by teachers. One such method is Tukey Tri-Split (Tukey, 1977). Such an approach to interpretation can strengthen transparent and rational interpretation. Additionally, it can reduce intuitive guided and teacher-dependent interpretations, as observed in the context of high-stakes decisions (Vanlommel and Schildkamp, 2019). Nevertheless, it can currently only be assumed that an instruction to perform Tukey Tri-Split actually increases the ability to predict future learning. Thus, one aim of our paper is to contribute to this desiderate.

Tukey Tri-Split: A non-arithmetic method for determining the slope of a learning progress graph

The Tukey Tri-Split (Tukey, 1977), also referred to as the Median Based Slope, is a graphical method by which a trend line can be plotted based on the first and third segments of a dot-line plot divided into three sections. This fairly simple-to-implement and non-arithmetic approach is widely used in school-based single case research (Vannest et al., 2013; Parker et al., 2014) and is also generally recommended when interpreting learning progress data for the purpose of making educational support decisions (Hosp et al., 2007; Fuchs and Fuchs, 2011). The basic idea behind this is that teachers can use this guided approach to determine the slope of learning development graphically and without numerical calculations. To do this, they proceed as follows:

1. The existing learning progress graph is divided into three equal-sized sections. If the number of measurement time points is not divisible by three, the division is made in such a way that the first and third segments are of equal length and the middle segment is the longest (see Figure 1, step 1).
2. The median is determined graphically for the first and third segments (see Figure 1, steps 2–3). To do this, the intersection point of the y-axis is exceeded and undershot by an identical number of points of the corresponding segment. The median of the segment

is marked in the middle of the segment in relation to the x-axis.

3. The trend line (slope) results from the connection of the two markings in the two segments (see Figure 1, step 4).

The trend line emerging from the tri-split can be used as a guide to estimate future learning development, provided that instruction is assumed to remain unchanged.

Present study

In this paper, we present two interrelated studies in the area of graph literacy. The overarching goal is to gain a detailed look into the process of interpreting learning progress graphs of inexperienced university student teachers in the first phase of teacher training in order to develop and evaluate a targeted intervention based on these findings for novices. The lowest level of graph literacy (i.e., read the data; Zeuch et al., 2017) will be given special focus in order to design a low-threshold intervention for this target group. Therefore, the sample considers student teachers and in-service teachers from Germany, a country without an implemented MTSS school system. From this combination, information can be derived to sharpen the content of the intervention for novices. In Study I, we ask which data base and which interpretation strategy student teachers use to predict a future learning progress depending on multiple graphs' characteristics. Consistent with the considerations about the state of the German school system, Study I (see section "Results" in this paper) showed that participants used a rather narrow and inconsistent data base to predict future learning developments. Thus, in Study II, student teachers and in-service teachers are trained to accurately predict learning progress with a 3-min video-based intervention. We adopt the graphs based on the results of Study I and create a short video tutorial about how to predict a further learning outcome relating on formative data. Research questions and methods are described for each study separately in the following sections.

Study I: Student teachers' approach to graph literacy

Research questions

Study I focuses on intuitive graph literacy by untrained student teachers. Previous research suggests that inexperienced teachers, on the one hand, do not have a consistent approach to reading formative data (Wagner et al., 2017; Blumenthal et al., 2021) and, on the other hand, that the structure of the data can influence the prediction of future progress (e.g., Keuning et al., 2017; Klapproth, 2018). However, it is unknown what specific

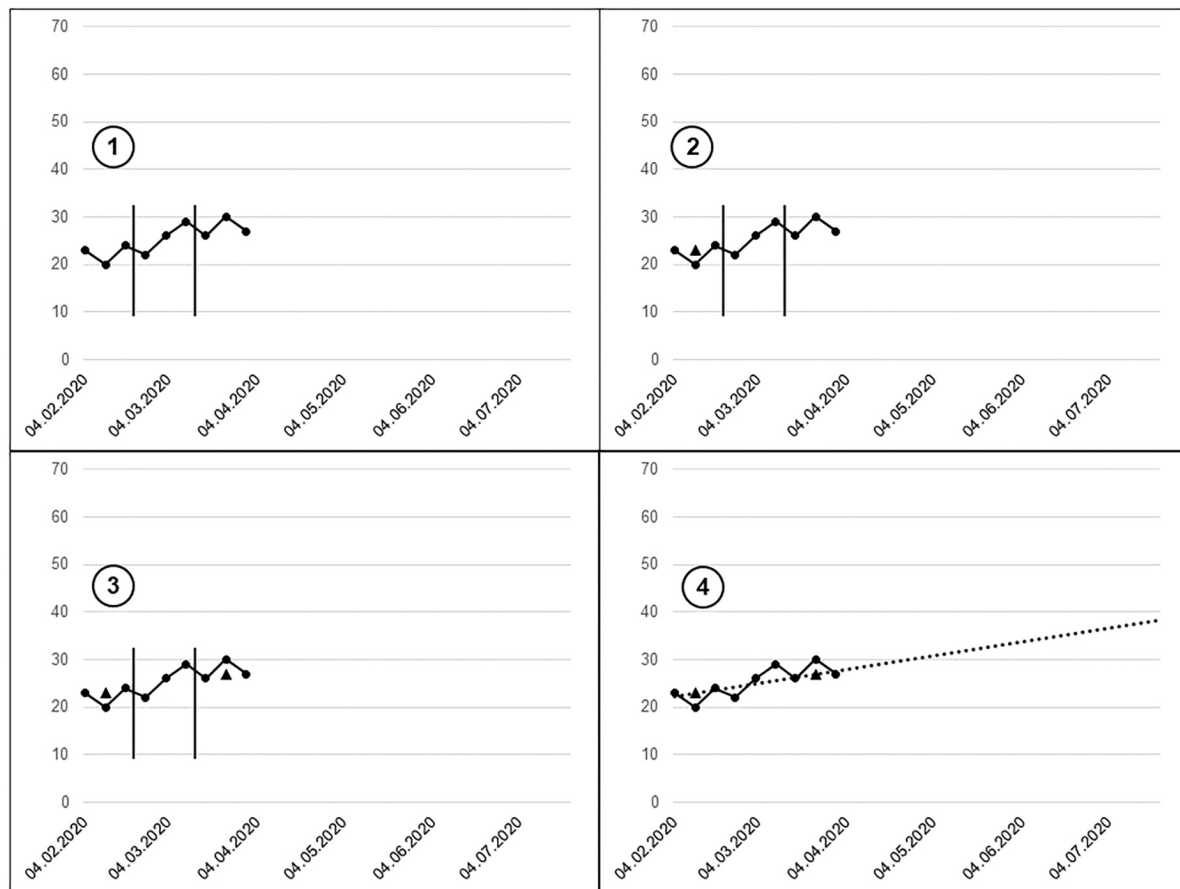


FIGURE 1
Demonstration of the Tukey Tri-Split (Scheer, 2021).

data they focus on and what strategy they intuitively use without specific instruction. Therefore, we ask three research questions:

1. How distinctive is prior knowledge of formative assessments and mathematical skills of student teachers?
2. Which approach regarding data base and interpretation strategy do untrained student teachers use to predict a future learning progress depending on multiple graphs' characteristics most often?
3. How stable are student teachers' decisions regarding their chosen data base and interpretation strategy across multiple graphs?

Methods

Sample and procedure

German student teachers enrolled in a primary, secondary or special school teacher education program were recruited *via*

social media platforms. The study was realized as a standardized web-based survey platform called limesurvey.org. The survey was online for 6 weeks.

In total, $N = 349$ student teachers participated. For data cleaning, participants who did not answer any question about the data prediction were removed. Thus, data from $N = 196$ participants from four German federal states [i.e., North Rhine-Westphalia (82.2%), Lower Saxony (14.8%), Saxony (1.5%), and Bremen (0.5%)] were analyzed. Most of the participants were female (82.1%), aged between 21 and 24 years (59.7%) and enrolled in the Bachelor's program (67.9%). They aimed to graduate in elementary school (20.9%), secondary school (27.6%) or special education school (50.5%) teacher programs.

Instrument

The web-based survey included a formal instruction, questions about background variables, four questions about prior knowledge regarding graph literacy, one example and six graphs (i.e., case vignettes). All case vignettes were presented to the participants in the same order and on the same screen with the questions about prediction and graph literacy.

Prior knowledge and skills

Participants were asked to self-assess their prior knowledge regarding (1) the approach of formative assessment and (2) graph reading in an educational context both with a four-point rating scale (responses ranged from [1] “no prior knowledge” to [4] “a large amount of both theoretical and practical application knowledge”). In addition, they were asked to assess their skills in (3) mathematical competencies and (4) graph reading in mathematical contexts (six-point rating scale, responses ranged from [1] “very good” to [6] “very bad”).

Case vignettes

Each case vignette displays a learning progress graph. All graphs were constructed and manipulated following the study material of Klapproth (2018). Figure 2 shows the first case vignette. The x-axis represented 14 school weeks as time line. At the y-axis, the number of correct read words per minute (WRC) were marked. The first eleven data points were given, which were separated into three graph sections: baseline including three data points, 1st intervention phase and 2nd intervention phase including each four data points. For each graph section, a separate linear trend line was presented. For this study, the graphical subdivisions and the addition of the trend lines were necessary to gain insight into the data base and strategies used. Additionally, the baseline of the peers and a theoretical maximum were given. All six graphs were based on the following linear function: $WRC = bx + a$ with b representing the slope, x the school week, and the intercept. The graphs were manipulated in two aspects. First, the graphs differ in a low, middle and high rate of improvement (i.e., b or $1.3*b$ or $3.4*b$). Second, the variation of the data points was either low or high (i.e., b or $2*b$). All experimental data points were calculated according to progress monitoring data of German second graders in reading (Anderson et al., 2020). Participants were asked to predict the data points for weeks 12 and 13 as numerical values based on the available data for each case vignette.

Graph literacy

In a closed-response and single-choice format, participants were asked for each case vignette which data base (Which data did you use for your prediction?) and which interpretation strategy (How did you predict the learning growth?) they used. The given answers were initially based on a preliminary exploratory study with special education student teachers, which clustered possible strategies by a content analysis according to Mayring (2014). In a second step, the clustered answers were cross-referenced with models of graph literacy (Zeuch et al., 2017) and with possible influencing variables that might condition errors in predicting a learning growth (Keuning et al., 2017; Newell and Christ, 2017). The following answers regarding the data base used were available for selection: (1) baseline,

(2) 1st intervention phase, (3) 2nd intervention phase, (4) both intervention phases, (5) baseline and both intervention phases, (6) other time period (i.e., outside the specified phases), and (7) no time period (i.e., single data points). In this context, the first four and seventh responses represent a disregard of important information because not all available data were considered for interpretation. For answer six, there was an opportunity to describe the self-selected time period in more detail. The strategy used was inquired with the following contents provided: (1) concrete data points, (2) trend line, (3) pattern of the learning growth, (4) general instruction assumptions, (5) guessed, and (6) other strategy. The first four responses represent the Zeuch et al. (2017) levels, with the first and second responses being assigned to the reading the data level. All responses to this question were formulated as complete sentences to inquire the priority course of action. Therefore, the first four answers were worded with the addition “mainly.” The guessing strategy was derived from the qualitative responses of the preliminary study. Participants had the opportunity to describe their other strategy in writing.

Data analysis

Self-assessment differences on the four variables in prior knowledge and mathematical skills were tested using multivariate ANOVA. All data on graph literacy (i.e., data base and strategy) as well as the numeric prediction values were analyzed descriptively. In addition, the number of different responses to the data base and interpretation strategy was counted. To examine whether a particular approach to interpretation was used as a function of case vignette characteristics, we tested the reliability of the agreement (i.e., Fleiss' κ ; Fleiss, 1971). Afterward, we analyzed descriptively the number of switches within the approaches. We summed up the results related to the number of the choices of the data base and strategies under the term stability in graph literacy.

Results

Prior knowledge and skills

All student teachers estimated their prior knowledge regarding formative assessment at an intermediate level, with the formative assessment approach being slightly more common than dealing with graphs in an educational context. Across all student teacher groups (i.e., primary level, secondary level, and special education needs), they reported a mean of 2.30 ($SD = 0.70$) for graph reading in an educational context and a mean of 2.41 ($SD = 0.84$) for knowledge about the approach of formative assessments, while 4 was the maximum value. Only 3.5% of all participants stated that they also had practical application knowledge. With respect to mathematical skills, a mean of 3.29 ($SD = 1.05$) was reported for mathematical competencies and a mean of 4.13 ($SD = 1.12$) for graph reading

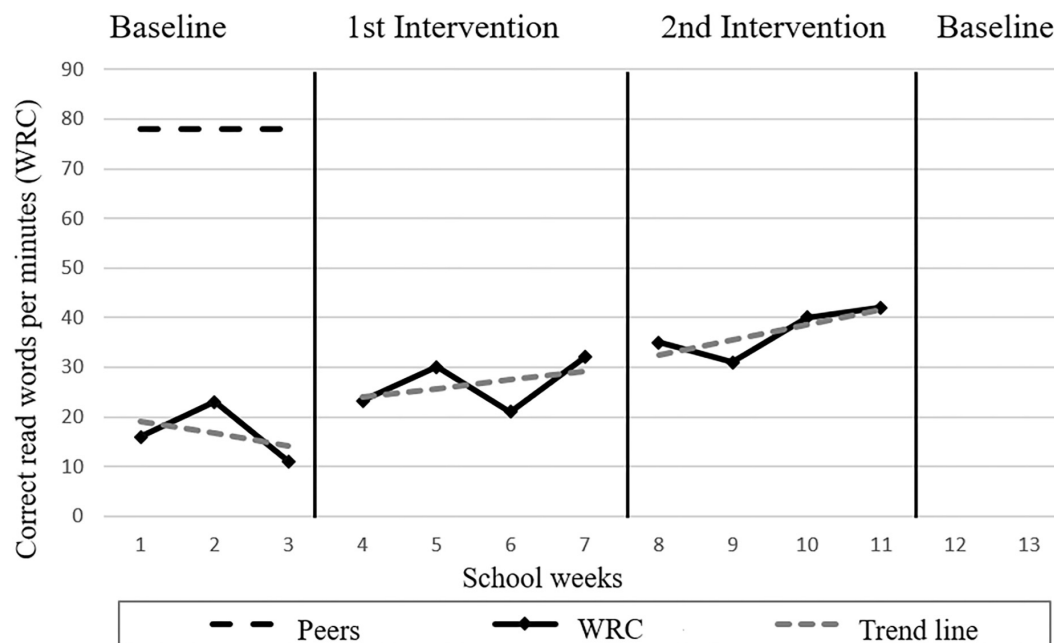


FIGURE 2
First case vignette in Study I.

skills in mathematical contexts, while the responses had a range from 1 to 6.

With regard to knowledge about the approach of formative assessments [$F(2, 191) = 0.766, p = 0.466$] and mathematical competencies [$F(2, 191) = 2.881, p = 0.59$], no significant differences were observed among the focuses of teacher training. The groups showed significantly different mean values with regard to knowledge of graph reading in an educational context [$F(2, 191) = 4.150, p > 0.05$] and in a mathematical context [$F(2, 191) = 4.572, p > 0.05$]. Tukey *post-hoc* tests showed that the student primary teachers rated their prior knowledge of graphs in an educational context significantly lower compared to the other teachers (compared to secondary: 0.36, 95%CI [0.03, 0.70], $p < 0.05$; compared to special education: 0.34, 95%CI [0.04, 0.64], $p < 0.05$) and that the special education student teachers rated their competencies lower than the secondary school student teachers in terms of mathematical competencies (−0.56, 95%CI [−1.00, −0.12], $p < 0.01$).

Approaches to graph literacy

Across all case vignettes, to predict future learning progress, student teachers most often considered data from the intervention phases: both intervention phases together (35.5%), only the second intervention phase (28.1%), all existing data points (i.e., baseline with both intervention phases, 19.7%), and only the first intervention phase (11.4%). Prediction based only on baseline (1.5%) or independent of any of the specified time periods (1.3%) were rarely reported.

As a strategy, they primarily used three approaches for prediction: continue the pattern of learning progress (41.0%), focus on the slope of the trend line (29.3%), and use other unspecified strategies (13.2%). The other three strategies were used similarly infrequently: assumptions about the instruction (7.0%), orientation on single measurement points (5.2%), and guessing (4.4%).

For a more detailed look, Figure 3 shows the absolute distribution of the selected strategies separated by the case vignettes and divided by the teacher programs. The arrangement of the case vignettes in Figure 3 is based on the 2×3 manipulation of the data (see also section Instrument of Study I). In the distribution of the selected strategies, per graph is sorted according to graph's slope (from top to bottom: low, medium, high) and in the columns according to graph's variation of the data points (left: high; right: low). The numbering of the graphs reflects the displayed order within the questionnaire. The distribution of strategies used per graph suggests that individuals switch their strategies when predicting. Moreover, there is no clear pattern in Figure 3 regarding the choice of strategy, which could be related to the characteristics of the graphs (i.e., rate of improvement and variability of the data).

Stability in graph literacy

Across all case vignettes, participants used a variety of data bases and strategies to make their predictions. The results show slight agreement for both data (Fleiss' $\kappa = 0.050; p < 0.001$) and strategy (Fleiss' $\kappa = 0.071; p < 0.001$). Considering the

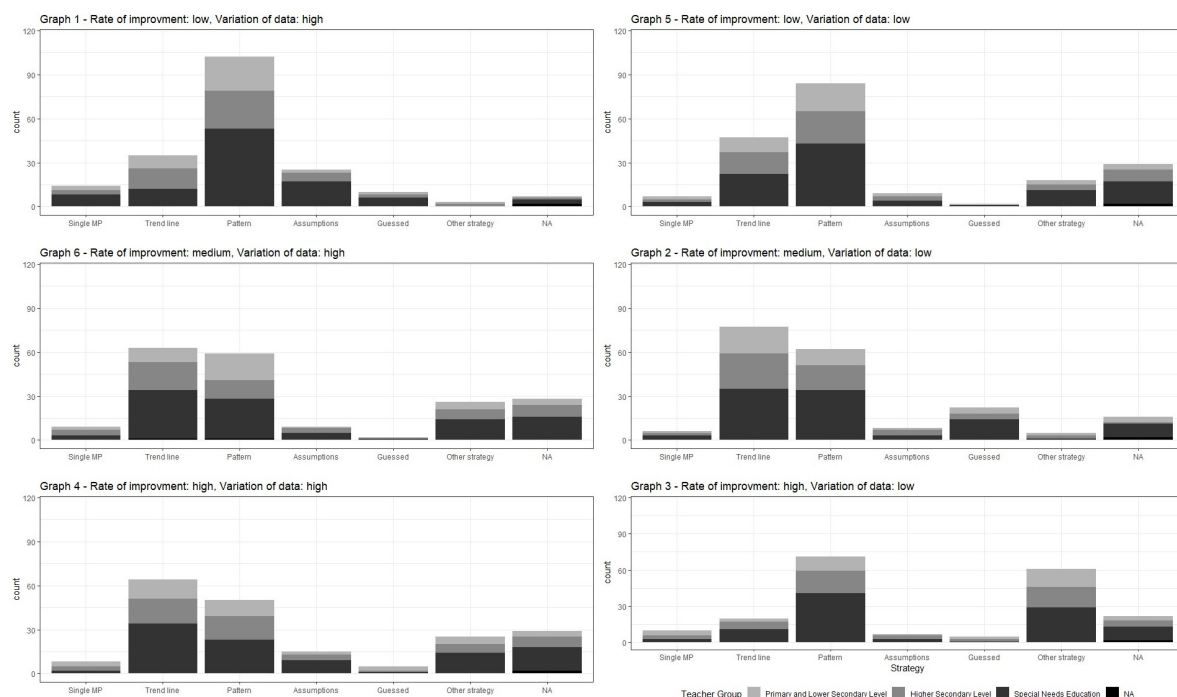


FIGURE 3
Absolute distribution of the selected strategies separated by the graphs and teacher program.

case vignettes individually, the levels of agreement differed significantly between participants' answers to graph literacy questions, ranging from -0.016 to 0.439 for data used and from 0.004 to 0.374 for strategies. Moderate agreement was found for the used data for graph 6 (Fleiss' $\kappa = 0.354 - 0.374$; $p < 0.001$) and fair agreement for the interpretation strategy for graphs 1 and 4 (Fleiss' $\kappa = 0.439$; $p < 0.001$). These results illustrate two things. First, the use of the data points is more coherent than that of the interpretation strategy. Second, student teachers do not have a consistent approach to prediction.

A switch within approaches to graph literacy could be observed in almost all participants. All student teachers used at least two different kinds of data bases to predict the further data points across the six case vignettes. 38.6% of the student teachers used two or three different data bases. Four different kinds of data bases were used by 21.7% of the student teachers. Only two persons used five different data bases (1.2%). In addition, changes in strategy were observed for all student teachers, except for one person. More than half of the student teachers used three different strategies (56.3%) across the six case vignettes. The remaining student teachers changed their strategy two (24.7%), four (13.9%), or five times (4.4%).

Discussion

Examination of the prior knowledge of the sample from Study I suggests that the approach to formative

assessment has been consistently weak among student teachers. While most student teachers are aware of the existence of formative assessments and proportionately have theoretical background knowledge, hardly any participants reported practical experience in their use ($< 5\%$). Graph reading experiences in educational and mathematical contexts differed significantly by teacher training. This is to be expected in the university teacher training in Germany, as the proportions of educational and subject-specific training contents are weighted differently depending on the field of study and individual focus in the teacher training program.

With regard to the choice of the data base for prediction, Study I shows that student teachers have a high risk of an unrealistic estimation of future learning progress. Only about 20% of the students intuitively included all available data in their prediction, which is, however, necessary for an accurate prediction (Espin et al., 2017; Klapproth, 2018). Over half of the students focused on a subset of the available data points rather than all available information. Thus, the predictions made about future learning were predominantly based on insufficient data.

Examination of the strategies chosen highlights that there is a great need for specific instruction on graph literacy because student teachers showed an inconsistent approach to prediction. No systematic reason for the choice of strategy can be identified in the available data, such as a property of the graph or a

preference by focus in study. Additionally, they frequently switched their strategy. Students most frequently used those strategies (i.e., continuing the pattern or orienting to the trend line) that fall into the two lower levels of graph literacy competence according to Zeuch et al. (2017). The results show that all participants except one switched their prediction strategy within the six case vignettes.

Study I is limited in multiple ways. First, we could not pre-determine the sample size and did not have a really representative randomized sample but an *ad hoc* sample of persons willing to participate in a survey on this specific subject. Thus, a potentially higher motivation compared to that of the average population of student teachers might bias the results. Teachers with average motivation might therefore show more severe or other difficulties in interpretation. Second, the numerical predictions could not be used to validate the selected data base and interpretive strategy due to the layout of the case vignettes. Visual aids were included in the layout of the graphs as possible factors influencing prediction such as trend line, division between baseline and intervention phases following previous research (Keuning et al., 2017; Newell and Christ, 2017) to provide a nuanced insight into the graph literacy approach. This ensured that even the most inexperienced student teachers could make statements about their prediction procedure. However, the embedding of visual aids means that the assumption about a linear trend in learning progress across all data points is not tenable. Thus, a reference value for matching the accuracy of prediction is missing. In addition, the fixed order of the case vignettes presented may have led to effects in prediction. This design was implemented based on the pilot study to avoid confusing very inexperienced student teachers at entry. In similar studies, such effects should be taken into account or eliminated by a randomized order.

Study II: Video-based intervention on graph reading accuracy

Research questions

The results from Study I suggest that student teachers tend to interpret learning progress graphs intuitively, without a systematic or consistent approach. However, especially for short- to medium-term prediction of future learning developments under the condition of unchanged teaching, it would be necessary to use information about the linear trend. A non-arithmetic approach to estimate the slope of the regression line is Tukey Tri-Split (Tukey, 1977). In Study II, we investigate whether brief video-based instruction on this method increases student teachers' and in-service teachers' short-term predictive accuracy on learning developments compared to a simple text-based hint to consider linear trends.

As van den Bosch et al. (2019) show, teachers' graph literacy skills can be improved *via* video instruction. However, they used a more general approach which results in a complete instruction on graph comprehension. In Study II, as progress monitoring is still an emerging field in the German school system, we take one step back and ask if the first level of graph literacy, namely predicting learning outcomes by identifying linear trends, can be improved by a short video intervention. Furthermore, it was our aim to examine whether a less than 5 min instruction is sufficient to achieve an improvement among teachers in the field.

Our main hypothesis is:

H₁: Student teachers and in-service teachers who receive a very short video instruction about how to use Tukey Tri-Split will improve their short-term predictive accuracy on learning developments more than those who only receive a text-based hint to consider linear trend in data.

Thus, our Null-Hypothesis to be rejected is:

H₀: There will be no difference in short-term predictive accuracy on learning developments between student teachers and in-service teachers who receive a very short video instruction about how to use Tukey Tri-Split and those who only receive a text-based hint to consider linear trend in data.

Methods

Sample and procedure

Using the online learning platforms of the authors' universities, mail contacts to other universities, mail contacts from in-service teacher training providers, and social media platforms, we invited student teachers and in-service teachers to participate in an online survey about learning progress monitoring. In total, $N = 198$ participants completed the survey.

Within this survey, we implemented a randomized controlled trial: At the beginning of the survey, which was implemented with the software Unipark, all participants received four case vignettes of Study I with the same prediction task estimating numerical values for 1 and 2 weeks after the last measurement point (i.e., weeks 12 and 13) as the pretest. After the pretest, about half of the participants ($n = 100$) were assigned to the experimental condition. They were shown a short instructional video, introducing Tukey Tri-Split and explaining it with an example. The other half ($n = 98$), as a control condition, received a text-based hint to consider linear trend in data. Finally, all participants completed the same prediction task with the same four case vignettes again as the posttest.

Conditions

During the survey, participants were randomly assigned to either the experimental or the control condition. The random

trigger variable in Unipark was set to provide a nearly equal distribution between both conditions.

Experimental condition

In the experimental group (EG), participants received a 03:03 min video instruction which introduces the Tukey Tri-Split method. The video script adopted the explanation from Hosp et al. (2007) in the way it was transferred to the German school context by Scheer (2021) and embedded it within the example of a primary school teacher wanting to decide which pupils need additional support in reading fluency. The video script and the video in German itself were provided *via* OSF (see section “Data Availability Statement”). To ensure that the given example in the video was different from the case vignettes, we used the example from Figure 1 as the basis for instruction.

Control condition

The procedure under control condition was the same as under experimental condition except for the intervention between pre- and posttest. Participants in the control group (CG) received, instead of the video, the following text hint:

“Very good. You have completed the first half. In the second half of the survey, we will show you the case vignettes again. Please consider the following tip: Ask yourself whether you can recognize a certain (linear) development trend in the available data, which you can use as a guide.”

We utilized this as a non-specific treatment component control instead of a no-treatment control (Mohr et al., 2009). The rationale for this decision was to ensure that systematic instruction of a specific technique was indeed necessary to improve prediction accuracy and that participants in the experimental group did not improve by priming on one specific feature of the history plots alone.

For ethical reasons, participants under control condition were offered the opportunity to watch the video instruction after submitting the survey.

Measurements

Predictive accuracy on learning progress

To reduce the burden on participants, only four (i.e., in the order presented: graph 6, graph 2, graph 4, and graph 3) of the six case vignettes were used in Study II. In all case vignettes, all optical aids (labeling baseline and intervention phases, vertical lines) or rate of improvement (slope) information were removed (for example, see Figure 4). Thus, it was possible to maintain the assumption of linear trend across all data points.

The graphs’ characteristics varied according to rate of improvement (medium vs. high) and variability of data points

(low vs. high). Thus, the four case vignettes represent a full 2×2 combination of both characteristics.

To calculate a score of prediction accuracy, we followed the approach of constant errors (CE) as used by Klapproth (2018). CE is calculated as the difference of a participant’s prediction of learning outcome (PP) and the learning outcome as predicted by regression (PR). PR was calculated using the arithmetic algorithm to replicate Tukey Tri-Split.

However, since we needed average test scores across individual case vignettes, we had to eliminate negative deviations by squaring CE, resulting in a Squared Constant Error (SCE). To achieve a total test score, we averaged SCE across all eight values (four case vignettes with two data points to be predicted each), resulting in a Mean Squared Constant Error (MSCE). Table 1 gives an overview of these measures. Squaring CE to SCE/MSCE also leads to a kind of penalty for more inaccurate PP compared to PP close to PR.

Treatment fidelity

To validate our results, we asked the participants under experimental condition to rate on a four-point scale:

1. Did you watch the explanatory video shown in the middle of the survey in full and fully concentrated on it?
2. Were you able to follow the explanations in the video well?
3. Were you able to apply the method presented in the video to the case studies that followed?

Furthermore, we asked the participants under control condition to rate on a four-point scale:

1. Was the hint (linear trend) in the middle of the survey helpful?
2. Did you change your approach after the hint?

Background variables

To examine if both the experimental and control groups were comparable with regard to their personal and professional background, we collected data on participants’ profession (special needs education teacher training vs. regular teacher training), gender, age as well as self-rated prior knowledge in learning progress monitoring, graph comprehension, and general mathematics skills.

Data analysis

We only included participants with correct participants code to ensure that no duplicates bias the analysis and with all case vignettes completed.

Since outliers are a serious source for bias, we applied the interquartile range (IQR) approach to detect any outliers. Thus,

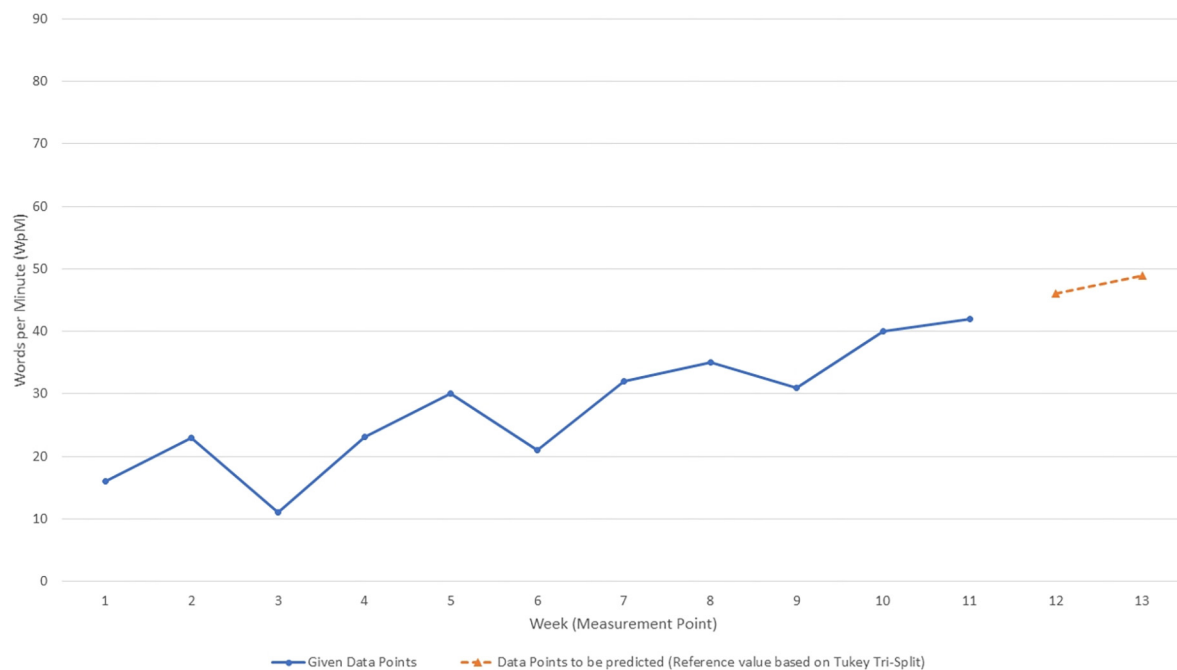


FIGURE 4

Example of a learning progress graph used in Study II. For this publication, we added the reference values for the data points that had to be predicted by the participants (orange). The y-axis represents week of learning progress measurement, the x-axis represents pupil's learning outcome (reading fluency, words per minute).

TABLE 1 Overview of constant error (CE), squared constant error (SCE), and mean squared constant error (MSCE), and their application in the study.

Abbrev.	Name	Description	Formula
CE	Constant error (Klapproth, 2018)	Difference between participant's prediction (PP) and predicted value from regression analysis (PR)	$CE = PP - PR$
SCE	Squared constant error	Square of the difference between participant's prediction and predicted value from regression analysis; used as test score per graph/data point	$SCE = (PP - PR)^2$
MSCE	MSCE	Mean of the SCE across all four case vignettes; used as total test score for pre- and posttest	$MSCE = \frac{\sum_{i=1}^k (PP_i - PR_i)^2}{k}$ with k items

participants were classified as outliers if one of their MSCE (post- or pretest) was either 1.5 times IQR above the third quartile (Q_{75}) or below the first quartile (Q_{25}). In the case of an online study with no control over participants' attention while answering the test items, outliers are considered as caused by inattention or typos when handling the online survey tool. Therefore, to avoid biased analysis, we excluded cases who were classified as outliers.

Using 2×2 ANOVA with a within-subject factor (pre- vs. posttest) and a between-subject factor (experimental vs. control), we tested whether the video intervention had a significant effect on the MSCE score.

An explorative follow-up analysis was performed to analyze whether graph characteristics (rate of improvement, data variability) and distance from last data point (namely: week 12 vs. week 13) have an impact on both the SCE and

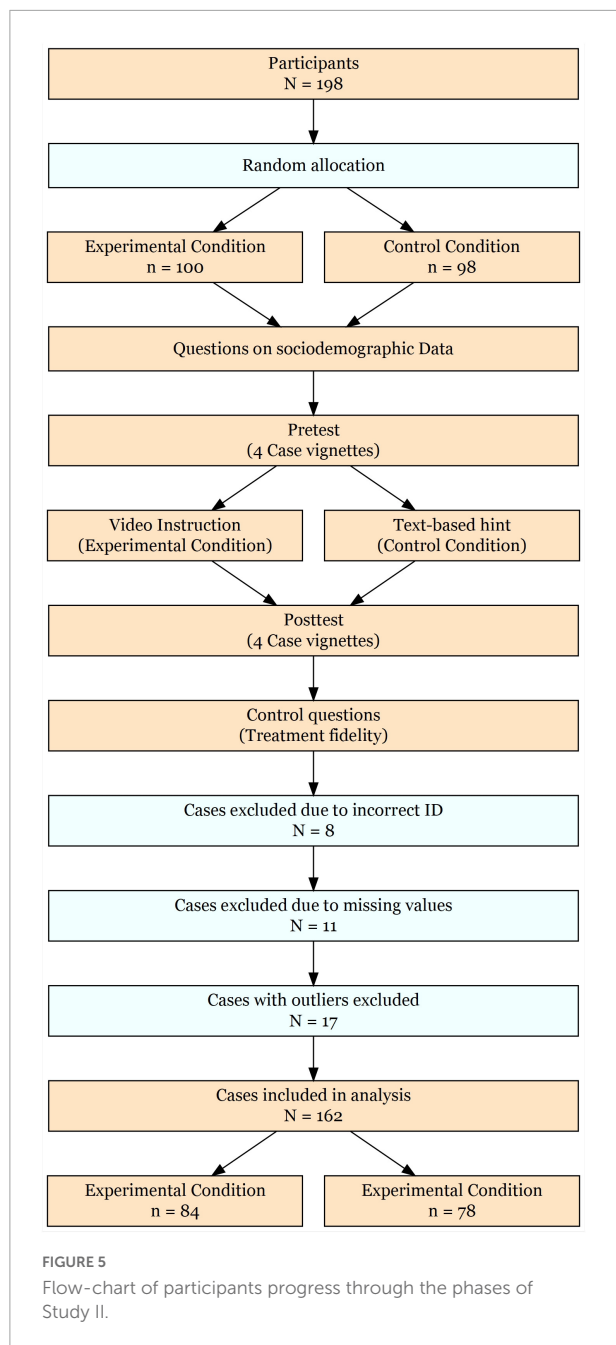
the intervention effect. To that purpose, we used stepwise linear regression.

Results

Sample characteristics

From $N = 198$ participants who completed the survey (EG: $n = 100$; CG: $n = 98$), eight participants (4.0%) were excluded due to incorrect user ID, eleven (5.6%) due to missing values, and 17 (8.6%) due to outliers. Thus, we analyzed a total sample of $N = 162$ participants with $n = 84$ in EG and $n = 78$ in CG (see Figure 5).

On average, the participants were 31.1 years old ($SD = 13.1$), with 78.9% ($n = 127$) being female (male: 20.5% [$n = 33$]; other gender: 0.6% [$n = 1$]). The majority of participants were



student teachers (56.8%), the dominant teaching degree across all participants was special needs education (80.1%, $n = 129$). Only a minority of the participants had had prior contact to progress monitoring (32.1%, $n = 52$) or general graph interpretation (43.1%, $n = 69$). As Table 2 shows, no significant differences between EG and CG could be found with regard to the background variables.

Global treatment effect

As displayed in Table 3 and Figure 6, MSCE in EG decreased after the video-based instruction from $M = 55.83$

($SD = 30.40$, 95%CI [49.33, 62.33]) to $M = 31.98$ ($SD = 25.53$) while MSCE in CG remained nearly the same. The 2×2 RM-ANOVA indicated significant main effects for group (EG vs. CG) with $F(1, 160) = 8.63$, $p = 0.004$, Cohen's $f = 0.36$, for measurement time (pre- vs. posttest) with $F(1, 160) = 24.21$, $p < 0.001$, Cohen's $f = 0.39$ as well as a significant interaction effect of group vs. measurement time with $F(1, 160) = 24.26$, $p < 0.001$, Cohen's $f = 0.39$. Thus, we most probably may discard the H_0 and assume that our video-based Tukey Tri-Split instruction significantly improved participants predictive accuracy compared to a text-based hint.

Treatment fidelity

Table 4 displays the results of the questions for treatment fidelity. Most of the participants from the EG answered the treatment fidelity questions at least with “rather yes,” but were more self-critical when it came to rating their personal ability to implement the Tukey Tri-Split. Furthermore, MSCE in posttest was more below MSCE in pretest for participants who answered the treatment fidelity questions more positive. As for the most important question, whether participants watched the video completely and with concentration, the difference between those participants who answered at least “rather yes” and the rest was indicated as statistical significant by a two sample t -test with $t(df = 82) = -2.56$, $p = 0.012$, $d = -0.95$ (see Table 4 for details). Since the MSCE quantifies the deviation from the calculated predicted value, this result means that the predictions of the participants who watched the intervention video with more concentration were closer to the calculated target value at the posttest than the predictions of those participants who did not watch the video intensively. Participants from the CG, on the other hand, did not find their text-based hint helpful and showed no clear patterns of intervention effect based on their responses to the treatment fidelity questions.

Effects of graph characteristics on predictive accuracy and on the effectiveness of the video-based instruction

A full table of SCE descriptive statistics by measurement time, group, and graph characteristics is provided as Electronical Supplement (Supplementary material) via OSF (see section “Data Availability Statement”). The most important findings are, as Figure 7 illustrates, that participants from the EG showed higher SCE scores in the pretest of all four case vignettes. This means that the EG participants predicted the target value more accurately in the posttest than in the pretest, regardless of graph characteristics. For the EG group, the intervention effect, based on visual inspection, was found to be the largest for the graph with high variability and medium rate of improvement. Here, the

TABLE 2 Sample characteristics in Study II.

Variable	Overall	Experimental group	Control group	<i>p</i>	SMD	Missing
N	162	84	78			
Age [<i>M</i> (<i>SD</i>)]	31.1 (13.1)	31.9 (13.5)	30.4 (12.9)	0.473	0.114	0.6
Gender (%)				0.594	0.165	0.6
Female	127 (78.9)	65 (77.4)	62 (80.5)			
Male	33 (20.5)	18 (21.4)	15 (19.5)			
Other gender	1 (0.6)	1 (1.2)	0 (0.0)			
Profession (%)				0.845	0.091	0.0
Student teacher	92 (56.8)	46 (54.8)	46 (59.0)			
In-service teacher	49 (30.2)	27 (32.1)	22 (28.2)			
Other	21 (13.0)	11 (13.1)	10 (12.8)			
Teaching degree = special needs education (%)	129 (80.1)	64 (76.2)	65 (84.4)	0.268	0.208	0.6
Prior contact to progress monitoring = No (%)	52 (32.1)	28 (33.3)	24 (30.8)	0.856	0.055	0.0
Prior contact to graph interpretation = No (%)	69 (43.1)	38 (45.2)	31 (40.8)	0.684	0.090	1.2
Current skills in progress monitoring [<i>M</i> (<i>SD</i>)]	2.4 (1.0)	2.3 (1.0)	2.4 (1.0)	0.376	0.140	0.0
Current skills in Maths [<i>M</i> (<i>SD</i>)]	3.2 (0.9)	3.3 (0.9)	3.1 (1.0)	0.218	0.194	0.0
Current skills in Mathematical graph interpretation [<i>M</i> (<i>SD</i>)]	3.0 (0.9)	3.0 (0.9)	3.0 (0.8)	0.787	0.043	0.0
Current skills in progress monitoring graph comprehension [<i>M</i> (<i>SD</i>)]	3.0 (0.9)	3.1 (0.9)	3.0 (1.0)	0.519	0.101	0.0

p represents the significance of differences between EG and CG. Group differences were tested using χ^2 -test with continuity correction for categorical variables and using t-test for continuous variables. SMD represents measures of standardized mean difference.

TABLE 3 Descriptive statistics of the MSCE scores by group and measurement time.

	<i>N</i>	Pre-test				Post-test			
		<i>M</i>	<i>SD</i>	<i>SE</i>	95%CI	<i>M</i>	<i>SD</i>	<i>SE</i>	95%CI
Experimental group	84	55.83	30.40	3.3169	[49.33, 62.33]	31.98	25.53	2.7854	[26.52, 37.44]
Control group	78	54.79	31.26	3.5397	[47.85, 61.73]	55.27	27.55	3.1193	[49.16, 61.38]
Total sample	162	55.33	30.73	2.4141	[50.6, 60.06]	43.19	28.90	2.2705	[38.74, 47.64]

SCE value decreases from 52.17 at pretest to 13.54 at posttest. Furthermore, for all participants from the EG and CG, low variability of data points combined with a medium rate of improvement led to such a predictive accuracy in the pretest that there seemed to be no further intervention effect.

To exploratively analyze the impact of the graph characteristics, we conducted stepwise linear regression in four steps:

- Model 1 is the baseline model which just replicates the original analysis of the treatment effect itself (predictors: group, measurement time).
- In Model 2, we included the rate of improvement (medium vs. high) as predictor.
- In Model 3, we included variability of data points (low vs. high) as predictor.
- In Model 4, we added the distance of the predicted data point to the last given data point (week 12 vs. week 13) as predictor.

We compared the four models with regard to R^2 , AIC, and BIC. The full regression table is provided as Electronic Supplement ([Supplementary material](#)) via OSF (see section “Data Availability Statement”). As displayed there, model 4 performed best ($R^2 = 0.23$, AIC = 28,460.2, BIC = 28,653.6). However, increased model fit from step three to step four is quite small. In-depth analysis shows that the general intervention effect is still there, even if controlled for graph characteristics. Furthermore, a high rate of improvement results in significant higher SCE, which represents a weaker predictive accuracy. Additionally, as seen in visual inspection, lower variability of data points results in smaller SCE leading to better predictive accuracy. However, this effect of low data variability is eliminated in the posttest unless participants are in CG or the graph has a high rate of improvement. A greater distance of the predicted measurement point from the last given measurement point was, counter-intuitively, associated with better performance regarding predictive accuracy except for EG in the posttest. For EG, this means that the video-based instruction worked so well that participants no longer performed worse in rating week 12 than in rating week 13.

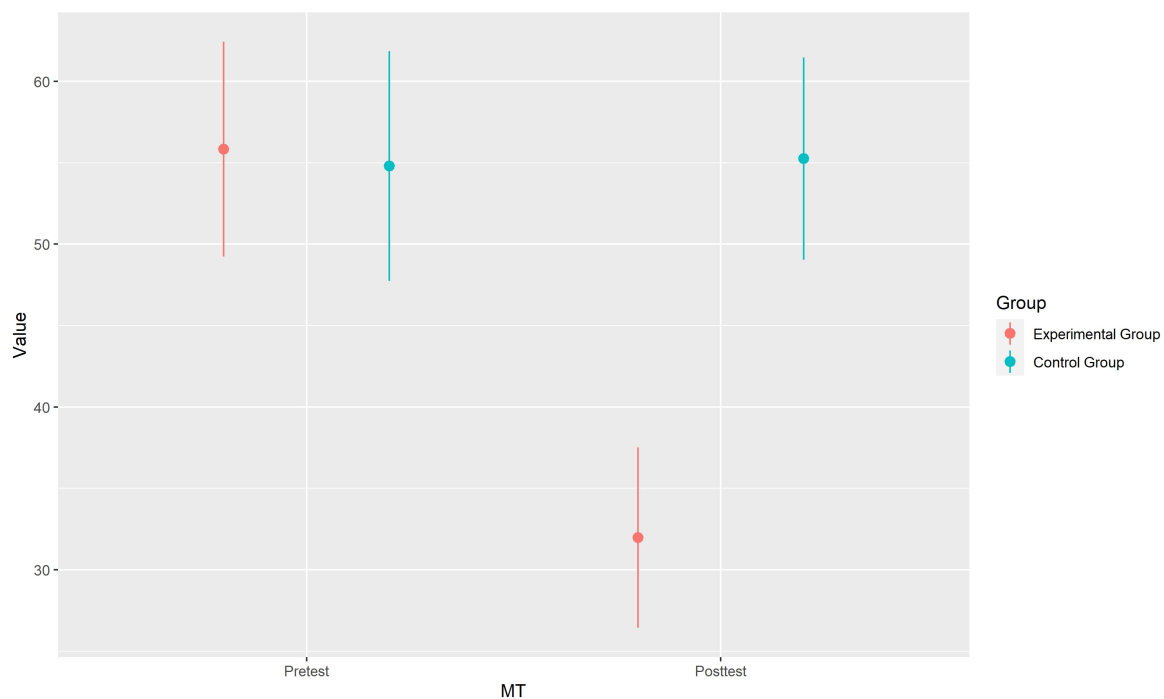


FIGURE 6
Interaction plot of MSCE scores by group vs. measurement time. Data points represent the group means. Error bars represent 95%CI.

Discussion

We could show that the video-based instruction of Tukey Tri-Split was effective in increasing student teachers' and in-service teachers' accuracy to predict pupils' future learning outcomes with a, according to Cohen (1988), large effect size of $f = 0.39$ compared to a simple text-based control group. The results of our treatment fidelity questions argue for a high amount of treatment fidelity as well as for the internal validity of the study. Participants who fully concentrated on the video and found it comprehensible did benefit more than those who did not. Our main question on treatment fidelity, however, contained both aspects, watching completely and with concentration. Despite this, some participants who watched the video completely but with less concentration might have answered "rather yes" or "rather no," although we assume that this presents only a small risk of bias.

Furthermore, we could show that participants' predictive accuracy was influenced by graph characteristics such as data variability (i.e., higher variability led to more inaccurate predictions), slope (i.e., higher rate of improvement led to more inaccurate ratings), and the week to be predicted (i.e., in pretest, week 13 was predicted more accurately than week 12). However, week 13 as point to be predicted and low data variability each reduced the effect of the video-based intervention, but did not eliminate it.

There are several limitations to be discussed. First, regarding the non-representative and non-randomized sample as well as

the motivation of the participants, the same difficulties show as in Study I. In both studies, this is due to the web-based realization of the questionnaires with voluntary participation. Second, predictions were made about graphs presented *via* computer display. If graphs had been available as printouts for the participants, effects might have been different. We do, however, assume that, in that case, the intervention effect might have been even higher—this is due to the fact that, in a pencil-paper-version, participants would have had the possibility to use rulers and draw on the diagram to make their predictions more accurate than when having to apply the technique on a computer screen. A third limitation follows from our control group: While the video intervention took 3 min, reading the textual hint in the control group might have taken just a few seconds. Therefore, we cannot preclude any effect of waiting time of any kind before post-test. A minor argument that could be included into the discussion are possible memory effects of the graphs. However, if such an effect had occurred, it should be the same for both groups, which was one reason to apply the randomized control design in our study.

General discussion

According to the U.S. Supreme Court decision in *Andrew F. v. Douglas County School District*, learning development and reaching support goals are the most important indicators to

TABLE 4 Participants' responses to the treatment fidelity questions and how these responses interact with the intervention effect.

		Yes	Rather Yes	Rather No	No	<i>t</i> (df)	<i>p</i>	<i>d</i>
Experimental group (EG)								
Did you watch the explanatory video in the middle of the survey completely and with concentration?	N	54	22	5	3	−2.56 (82)	0.012	−0.95
	Difference Pre-Post [Mean (SD)]	−24.9 (35.0)	−32.4 (40.7)	9.1 (28.9)	3.1 (22.2)			
Were you able to follow the explanations in the video well?	N	41	35	5	3	−1.56 (82)	0.123	−0.58
	Difference Pre-post [Mean (SD)]	−24.7 (28.8)	−27.2 (45.4)	−9.3 (36.0)	3.1 (22.2)			
Were you able to apply the method presented in the video to the case vignettes that followed?	N	15	38	28	3	−1.14 (82)	0.256	−0.26
	Difference Pre-Post [Mean (SD)]	−35.3 (36.3)	−24.2 (35.9)	−20.1 (39.2)	3.1 (22.2)			
Control group (CG)								
Was the hint (linear trend) in the middle of the survey helpful?	N	5	27	29	17	−0.65 (76)	0.520	−0.15
	Difference Pre-Post [Mean (SD)]	18.7 (14.8)	−5.4 (27.1)	0.9 (21.6)	3.7 (23.5)			
Did you change your approach after the hint?	N	4	13	38	22	0.28 (75)	0.781	0.08
	Difference Pre-Post [Mean (SD)]	−18.1 (62.7)	8.1 (25.9)	0.5 (18.0)	−0.6 (22.5)			

Difference Pre-Post is the difference between pretest MSCE and posttest MSCE on a subject level. Mean and SD are calculated on a group level. A two sample t-test was used to compare the mean difference in pre-post-difference. For conducting the t-test, groups has been collapsed by "Yes/Rather Yes" and "No/Rather No". *p* indicates the level of significance, *d* represents Cohen's *d*.

determine whether the chosen education is appropriate (Prince et al., 2018). Regardless of the school system, graph literacy is an increasingly important aspect of DBDM in inclusive and special education. Currently, benchmarks and goals for all students are often used as a standard of comparison. However, more important is the question of what learning development the individual student can achieve in his or her particular circumstances and what intervention is the optimal one. For such educational decisions based on quantitative progress monitoring data, simple tools such as the Tukey Tri-Split are necessary for teachers to define achievable learning goals (Hosp et al., 2007; Fuchs and Fuchs, 2011). A core competency of special education teachers is the goal setting and prediction of which goal will be achieved by the child. They must always consider under what conditions and in what environment the child learns best. How this competency can be improved in the area of assessment and graph reading for students and practitioners is an open question so far (Wagner et al., 2017; Blumenthal et al., 2021).

Our research focused on the lowest level of graph literacy (i.e., reading the data; Zeuch et al., 2017) in Study I in order to be able to develop a low-threshold intervention for novices in Study II. The results of Study I again replicate the need for specific support in graph literacy through an example with a

sample from Germany, a country without implemented MMTS. Even though few student teachers already intuitively take a good approach to predicting future learning progress from a relevant data base, this combination is so far rare and not consolidated. The approach of formative assessment originated in special education (Fuchs, 2017) and is also heavily researched and taught in Germany by representatives of this discipline (e.g., Jungjohann et al., 2018b; Blumenthal et al., 2021). It was surprising, therefore, that special education student teachers indicated equal amounts of prior knowledge and experience with formative assessments as did students in regular education. This finding suggests that it is not only graph literacy training that should be deepened, but also that awareness of the DBDM approach needs to be more widely disseminated across both teaching majors.

In Study II, we could show that a video-based instruction can increase student teachers' and in-service teachers' predictive accuracy of learning outcomes. Although our measurements are near-to-instruction measures, the findings are in line with other research (van den Bosch et al., 2019). However, since our instructional video was far shorter than those used by van den Bosch et al. (2019) with about 3-min against to up to 45 min, we could show that even very small and low-threshold interventions can have a huge impact, at least as a short-term

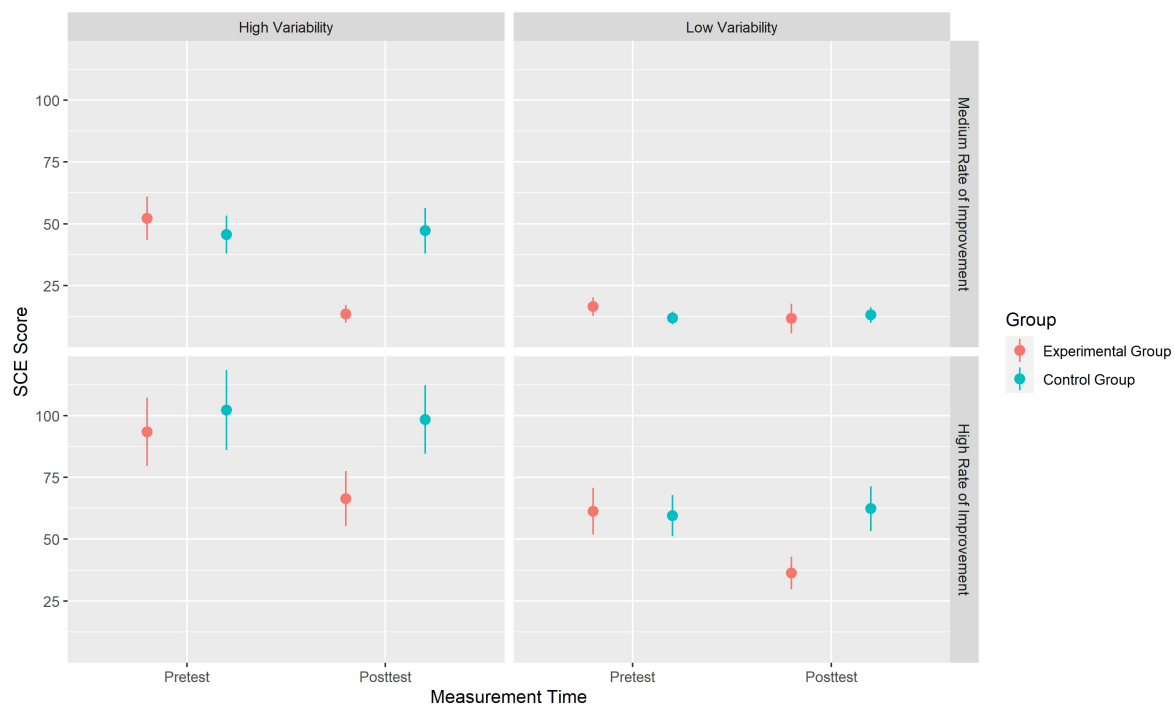


FIGURE 7

Interaction plot of the effectiveness of the video-based instruction by graph characteristics (rate of improvement vs. variability of data). Data points represent the group means and error bars represent the 95%CI. Figures in the left panels display results for the vignettes with high variability of data points while figures in the right panel display results for the vignettes with low variability of data points. The upper row displays the results for the vignettes with a medium rate of improvement while the bottom row displays the results for the vignettes with a high rate of improvement.

effect. In future research, it should be evaluated if there is (a) a transfer effect on DBDM skills in general and (b) a medium or even long-term effect.

Furthermore, Study II covered the research desiderate from Study I: We succeeded in evaluating whether graph characteristics (rate of improvement, data variability, distance from last point given) have any effect on predictive accuracy. Expectedly, graphs with a high rate of improvement and a high data variability were more difficult to interpret for the participants without training. Consequently, for these graphs, the video-based instruction had the biggest effect. For two reasons, this finding underpins the necessity of systematic instruction by using strategies for data prediction as for instance Tukey Tri-Split. First, in real-life learning progress monitoring, high data variability is expectable and, second, we want our interventions to increase the slope of learning progress. However, we need to further explore how accurate medium- to long-term prediction (for example predicting week 22 instead of week 12 when there are still 11 data points ahead) is and how instruction affects accuracy for these long-term predictions.

Reading graphs is an important component of DBDM (Mandinach and Gummer, 2016). However, this is only one component among many others. It is equally important to

interpret the other quantitative data from progress monitoring, in addition to the tasks solved, and to relate it to the other qualitative and quantitative data about the child and the learning environment. For comprehensive support, all data must be interpreted together as a team. Direct implications for school practice become apparent only when the entire process of DBDM is put into practice. Thus, in addition to school achievement tests and screenings, progress monitoring tests should be known and used in school practice. At present, this is not yet foreseeable in Germany for the next few years.

Limitations

One limitation across both studies concerns the transferability of the findings to school practice. In particular, Study II demonstrates the positive effects of the video-based intervention in terms of predictive accuracy. To what extent this improved prediction of short-term data has implications for the processes of DBDM in school practice remains to be seen. This will require, for example, a long-term study in the field focusing on prediction accuracy among teachers of their students. It would need to be verified whether the positive effects can also be replicated under the influence of other variables from

the field such as relation to teaching, interventions actually implemented, or individual learning paths.

Conclusion

Overall, we can conclude that there is a fundamental need to implement graph literacy skills into teacher training curricula for both general and special needs education. Such training can be integrated into existing teacher education. A few learning units on the central aspects of graph interpretation could be taught. These include the Tukey Tri-Split used in our study as well as the following topics: making conscious decisions about the number of measurement points, identifying the current state of learning distinguishing between baseline and intervention phases, and, last but not least, defining, setting and reviewing support goals. We can see that student teachers' and in-service teachers, without further training, lack strategies to interpret learning progress graphs. Our results furthermore indicate that even small but structured, direct-instructional training sessions such as the one used in our study can lead to important increases in graph literacy skills.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://doi.org/10.17605/OSF.IO/X2RS3>.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

References

- Ahmed, Y. A. (2019). Data-based decision making in primary schools in Ethiopia. *J. Prof. Cap. Commun.* 4, 232–259. doi: 10.1108/JPC-11-2018-0031
- Anderson, S., Jungjohann, J., and Gebhardt, M. (2020). Effects of using curriculum-based measurement (CBM) for progress monitoring in reading and an additive reading instruction in second classes. *Z. G.* 13, 151–166. doi: 10.1007/s42278-019-00072-5
- Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., and Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding curriculum-based measurement of oral reading fluency (CBM-R) decision rules. *J. Sch. Psychol.* 51, 1–18. doi: 10.1016/j.jsp.2012.09.004
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assess. Educ.* 18, 5–25. doi: 10.1080/0969594X.2010.513678
- Blumenthal, S., Blumenthal, Y., Lembke, E. S., Powell, S. R., Schultze-Petzold, P., and Thomas, E. R. (2021). Educator perspectives on data-based decision making in Germany and the United States. *J. Learn. Disabil.* 54, 284–299. doi: 10.1177/0022219420986120

Author contributions

JJ and MG designed the Study I and conducted the data collection. DS, JJ, and MG designed the Study II and conducted the data collection, designed the data analysis strategy for Study I, designed the data analysis strategy for Study II, and wrote and edited the manuscript. JJ wrote the R syntax for Study I and outlined the structure of the manuscript. DS wrote the R syntax for Study II. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by the Deutsche Forschungsgemeinschaft (DFG) (grant no. 453372524).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2022.919152/full#supplementary-material>

- Boy, B., Bucher, H.-J., and Christ, K. (2020). Audiovisual science communication on TV and YouTube: How recipients understand and evaluate science videos. *Front. Commun.* 5:608620. doi: 10.3389/fcomm.2020.608620
- Carlson, D., Borman, G. D., and Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educ. Eval. Policy Anal.* 33, 378–398. doi: 10.3102/0162373711412765
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd Edn. Mahwah, NJ: Lawrence Erlbaum Associates.
- Espin, C. A., Förster, N., and Mol, S. E. (2021). International perspectives on understanding and improving teachers' data-based instruction and decision making: Introduction to the special series. *J. Learn. Disabil.* 54, 239–242. doi: 10.1177/00222194211017531
- Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., and de Rooij, M. (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learn. Disabil. Res. Pract.* 32, 8–21. doi: 10.1111/ldrp.12123
- Fien, H., Chard, D. J., and Baker, S. K. (2021). Can the evidence revolution and multi-tiered systems of support improve education equity and reading achievement? *Read. Res. Q.* 56, S105–S118. doi: 10.1002/rq.391
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382. doi: 10.1037/h0031619
- Fuchs, L. S. (2017). Curriculum-based measurement as the emerging alternative: Three decades later. *Learn. Disabil. Res. Pract.* 32, 5–7. doi: 10.1111/ldrp.12127
- Fuchs, L. S., and Fuchs, D. (2001). *What is scientifically-based research on progress monitoring?*. Washington, DC: National Center on Student Progress Monitoring.
- Fuchs, L. S., and Fuchs, D. (2011). *Using CBM for progress monitoring in reading*. Washington, DC: U.S. Office of Special Education Programs.
- Gesel, S. A., Lejeune, L. M., Chow, J. C., Sinclair, A. C., and Lemons, C. J. (2021). A meta-analysis of the impact of professional development on teachers' knowledge, skill, and self-efficacy in data-based decision-making. *J. Learn. Disabil.* 54, 269–283. doi: 10.1177/0022219420970196
- Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Stud. Sci. Educ.* 47, 183–210. doi: 10.1080/03057267.2011.605307
- Gleason, P., Crissey, S., Chojnacki, G., Zukiedwicz, M., Silva, T., Costelloe, S., et al. (2019). *Evaluation of support for using student data to inform teachers' instruction (NCEE 2019-4008)*. Jessup, MD: National Center for Education Evaluation and Regional Assistance.
- Good, R., and Jefferson, G. (1998). "Contemporary perspectives on curriculum-based measurement validity," in *The Guilford school practitioner series. Advanced applications of curriculum-based measurement*, ed. M. R. Shinn (New York, NY: Guilford Press), 61–88.
- Hosp, M. K., Hosp, J. L., and Howell, K. W. (2007). *The ABC's of CBM: A practical guide to curriculum-based measurement*, 1st Edn. New York, NY: The Guilford Press.
- Jungjohann, J., DeVries, J. M., Gebhardt, M., and Mühling, A. (2018a). "Levumi: A web-based curriculum-based measurement to monitor learning progress in inclusive classrooms," in *Computers helping people with special needs. ICCHP 2018. Lecture notes in computer science*, eds K. Miesenberger and G. Kouroupetroglou (Cham: Springer International Publishing), 369–378. doi: 10.1007/978-3-319-94277-3_58
- Jungjohann, J., DeVries, J. M., Mühling, A., and Gebhardt, M. (2018b). Using theory-based test construction to develop a new curriculum-based measurement for sentence reading comprehension. *Front. Educ.* 3:115. doi: 10.3389/feduc.2018.00115
- Jungjohann, J., Diehl, K., Mühling, A., and Gebhardt, M. (2018c). Graphen der lernverlaufsdiagnostik interpretieren und anwenden – leseförderung mit der onlineverlaufsdiagnostik levumi [Interpret and apply graphs of learning progression monitoring - Reading support with online progress monitoring Levumi]. *Forsch. Spr.* 6, 84–91. doi: 10.17877/DE290R-19806
- Keuning, T., van Geel, M., and Visscher, A. (2017). Why a data-based decision-making intervention works in some schools and not in others. *Learn. Disabil. Res. Pract.* 32, 32–45. doi: 10.1111/ldrp.12124
- Klapproth, F. (2018). Biased predictions of students' future achievement: An experimental study on pre-service teachers' interpretation of curriculum-based measurement graphs. *Stud. Educ. Eval.* 59, 67–75. doi: 10.1016/j.stueduc.2018.03.004
- Kosslyn, S. M. (2006). *Graph design for the eye and mind*. Oxford: Oxford University Press.
- Kubinger, K. D. (2005). Psychological test calibration using the rasch model - some critical suggestions on traditional approaches. *Int. J. Test.* 5, 377–394. doi: 10.1207/s15327574ijt0504_3
- Lane, K. L., Oakes, W. P., Ennis, R. P., and Hirsch, S. E. (2014). Identifying students for secondary and tertiary prevention efforts: How do we determine which students have tier 2 and tier 3 needs? *Prev. Sch. Fail.* 58, 171–182. doi: 10.1080/1045988X.2014.895573
- Mandinach, E. B., and Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teach. Teach. Educ.* 60, 366–376. doi: 10.1016/j.tate.2016.07.011
- Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. Available Online at: <https://www.ssoar.info/ssoar/handle/document/39517> (accessed May 30, 2022).
- McMaster, K. L., Lembke, E. S., Shin, J., Poch, A. L., Smith, R. A., Jung, P.-G., et al. (2020). Supporting teachers' use of data-based instruction to improve students' early writing skills. *J. Educ. Psychol.* 112, 1–21. doi: 10.1037/edu0000358
- Mohr, D. C., Spring, B., Freedland, K. E., Beckner, V., Aream, P., Hollon, S. D., et al. (2009). The selection and design of control conditions for randomized controlled trials of psychological interventions. *Psychosom.* 78, 275–284. doi: 10.1159/000228248
- Newell, K. W., and Christ, T. J. (2017). Novice interpretations of progress monitoring graphs: Extreme values and graphical aids. *Assess. Eff. Interv.* 42, 224–236. doi: 10.1177/1534508417694855
- Okan, Y., Garcia-Retamero, R., Cokely, E. T., and Maldonado, A. (2012). Individual differences in graph literacy: Overcoming denominator neglect in risk comprehension. *J. Behav. Decis. Mak.* 25, 390–401. doi: 10.1002/bdm.751
- Oslund, E. L., Elleman, A. M., and Wallace, K. (2021). Factors related to data-based decision-making: Examining experience, professional development, and the mediating effect of confidence on teacher graph literacy. *J. Learn. Disabil.* 54, 243–255. doi: 10.1177/0022219420972187
- Parker, R. I., Vannest, K. J., and Davis, J. L. (2014). A simple method to control positive baseline trend within data nonoverlap. *J. Spec. Educ.* 48, 79–91. doi: 10.1177/0022466912456430
- Prince, A. M. T., Yell, M. L., and Katsiyannis, A. (2018). Endrew F. v. Douglas county school district (2017): The U.S. Supreme court and special education. *Interv. Sch. Clin.* 53, 321–324. doi: 10.1177/1053451217736867
- Scheer, D. (2021). *Toolbox diagnostics: Aids for (special) education practice [toolbox diagnostik: Hilfen für die (sonder-)pädagogische praxis]*, 1st Edn. Stuttgart: Kohlhammer Verlag.
- Schurig, M., Jungjohann, J., and Gebhardt, M. (2021). Minimization of a short computer-based test in reading. *Front. Educ.* 6:684595. doi: 10.3389/feduc.2021.684595
- Stecker, P. M., Lembke, E. S., and Foegen, A. (2008). Using progress-monitoring data to improve instructional decision making. *Prev. Sch. Fail.* 52, 48–58. doi: 10.3200/PSFL.52.2.48-58
- Tukey, J. W. (1977). *Exploratory data analysis. Addison-Wesley series in behavioral science quantitative methods*. Boston, MA: Addison-Wesley.
- van den Bosch, R. M., Espin, C. A., Pat-El, R. J., and Saab, N. (2019). Improving teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *J. Learn. Disabil.* 52, 413–427. doi: 10.1177/0022219419856013
- Vanlommel, K., and Schildkamp, K. (2019). How do teachers make sense of data in the context of high-stakes decision making? *Am. Educ. Res. J.* 56, 792–821. doi: 10.3102/0002831218803891
- Vannest, K. J., Davis, J. L., and Parker, R. I. (2013). *Single case research in schools: Practical guidelines for school-based professionals*. Abingdon: Routledge.
- Wagner, D. L., Hammerschmidt-Snidarich, S. M., Espin, C. A., Seifert, K., and McMaster, K. L. (2017). Pre-service teachers' interpretation of CBM progress monitoring data. *Learn. Disabil. Res. Pract.* 32, 22–31. doi: 10.1111/ldrp.12125
- Zeuch, N., Förster, N., and Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learn. Disabil. Res. Pract.* 32, 61–70. doi: 10.1111/ldrp.12126



OPEN ACCESS

EDITED BY

Erica Lembke,
University of Missouri, United States

REVIEWED BY

Michael Schurig,
Technical University Dortmund,
Germany
Kaitlin Bundock,
Utah State University, United States

*CORRESPONDENCE

Anne Foegen
afoegen@iastate.edu

†These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

RECEIVED 15 May 2022

ACCEPTED 22 August 2022

PUBLISHED 15 September 2022

CITATION

Stecker PM and Foegen A (2022)
Developing an online system
to support algebra progress
monitoring: Teacher use and
feedback.
Front. Educ. 7:944836.
doi: 10.3389/feduc.2022.944836

COPYRIGHT

© 2022 Stecker and Foegen. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Developing an online system to support algebra progress monitoring: Teacher use and feedback

Pamela M. Stecker^{1†} and Anne Foegen^{2*†}

¹Department of Education and Human Development, Clemson University, Clemson, SC, United States, ²School of Education, Iowa State University, Ames, IA, United States

A brief overview of the development of an online system to support algebra progress monitoring across several years of an iterative process of development, feedback, and revision is provided. Online instructional modules addressed progress monitoring concepts and procedures; administration and scoring of three types of algebra measures, including teacher accuracy with scoring; and navigation and use of the online data management system, including data entry, graphing, and skills analyses. In the final year of this federally funded research project, a test of the functionality of the completed system as well as an evaluation of teachers' knowledge, accuracy, and satisfaction with the online professional development was evaluated. Specifically, 29 general and special education secondary school teachers completed 11 fully developed online instructional modules independently and administered weekly two of three types of algebra measures across 10 weeks with one of their classes of students. Data analysis included teacher accuracy in the scoring of the measures; change in their knowledge of student progress monitoring and data-based decision making; and teacher satisfaction with the online system, including instructional content, feasibility, and usability for data-based decision making. Directions for future research and implications for classroom use of this online system are discussed.

KEYWORDS

professional development for teachers, progress monitoring, algebra, curriculum-based measurement, online learning, data-based decision making

Introduction

Progress monitoring is an essential component of data-based decision making (Espin et al., 2017). Progress data help teachers to pinpoint students throughout the year whose response to their mathematics program appears insufficient to meet year-end instructional benchmarks or goals. Research corroborates that teachers who use

progress monitoring to make instructional decisions, that is, teachers who revise student instruction when their data reveal inadequate progress, effect greater achievement than among students whose teachers use their own methods of assessment (Stecker et al., 2005, for review). For teachers to use data-based decision making effectively, however, they need to be knowledgeable users of technically sound progress data. Espin et al. (2017, 2021a), Wagner et al. (2017) demonstrated that teachers have difficulty, however, in using progress data for instructional decision making. Moreover, instructional supports, such as graphs with prompts about applying decision-making rules, student skills profiles illustrating levels of mastery by problem types, and consultation (in person or system-generated recommendations) may be needed to support teachers' effective use of data (Stecker et al., 2005; Jung et al., 2018; Fuchs et al., 2021). Professional development (PD) materials may include information, directions, and examples to support teachers' and preservice teachers' knowledge and skill acquisition in a particular domain. Espin et al. (2021b) examined available PD materials related to progress monitoring using curriculum-based measurement and coded content in four areas: general information, conducting progress monitoring, data-based decision making, and other. They found that data-based decision making was not addressed as much as the other topics and recommended that greater consideration be devoted to this area in future PD materials. The current PD project focused on progress monitoring in algebra. This fully online PD included general content about progress monitoring, information about conducting progress monitoring in algebra, and several features related to data-based decision making.

Although several conceptually based measures exist for algebra readiness (e.g., see Helwig et al., 2002; Ketterlin-Geller et al., 2015 for sample items and description), few technically sound measures are available for secondary mathematics in algebra. Foegen et al. (2017) have developed and established the technical adequacy of three types of progress monitoring measures for algebra (Espin et al., 2018; Genareo et al., 2019). Like learning rules and applying decisions for scoring some of the elementary-level reading (e.g., knowing types of miscues that count as errors in oral reading) and mathematics measures (e.g., scoring digits correct in answers), the content and scoring of the algebra measures requires explicit instruction to ensure accuracy, or reliability, of scoring and fidelity of implementation. For example, with the algebra measures, students construct written responses, and teachers score written papers, making judgments about whether answers are mathematically equivalent. One type of measure requires examination of student work on the item solution to determine whether partial credit should be awarded if the final answer is incorrect, but part of the solution is appropriate for reaching a correct answer. Because of teacher judgment involved in progress monitoring, accuracy in scoring and fidelity of implementation are critical for effective data-based

decision making. In response to interest in the measures, a professional development (PD) workshop was created in 2008 for practitioners and delivered in-person, most often with the PD staff going to the practitioners. While the in-person PD option increased access to the algebra progress monitoring measures, it was not feasible or cost-effective for individual teachers or for small districts, including those in more remote areas. The Professional Development for Algebra Progress Monitoring project was funded to address this need (Foegen and Stecker, 2009-2012). Over the course of 5 years, the research team worked with secondary teachers to develop, revise, and test an online PD system to make algebra progress monitoring accessible and efficient. In this paper, we describe briefly the development and features of the online system and the research results during the final year of the project on teachers' learning and their use of the system. Specifically, we examined whether teachers (a) could learn critical content about algebra progress monitoring from the online professional development and (b) be able to use the online system accurately and efficiently. Researchers also examined teacher satisfaction data about the system's content, navigation, feasibility, and usability.

Materials and methods

Algebra progress monitoring measures

The PD online system was developed to support three algebra progress monitoring measures (Algebra Basic Skills, Algebra Foundations, and Algebra Content Analysis) that had been developed during a previously funded project, Algebra Instruction and Assessment: Meeting Standards (AAIMS; Foegen, 2004-2007). During the earlier AAIMS grant, an iterative development process that incorporated teacher input, student data collection, statistical analyses to examine technical adequacy, and teacher feedback on the results was used over 4 years to refine and test five alternative algebra measures designed to reflect Pre-Algebra and Algebra 1 content typically addressed in grades 7–12. Based on our design and technical adequacy criteria, three of the five types of algebra measures were deemed acceptable for dissemination (Foegen et al., 2017; Espin et al., 2018; Genareo et al., 2019). Each of these three AAIMS measures had been based upon principles of curriculum-based measurement (Deno, 1985) that incorporated use of alternate forms of systematic sampling of core algebra skills or problem types that related to success in algebra, along with standardized administration and scoring procedures. Assessments were completed as relatively brief, timed paper/pencil tasks, either individually or as a whole class. Teachers scored the measures using the same scoring guidelines implemented in the research that established evidence of technical adequacy. Twelve parallel forms were developed for each of the three types of AAIMS measures. The three measures:

Algebra Basic Skills, Algebra Foundation, and Algebra Content Analysis, originally developed through the AAIMS grant, became the foundation for the current PD project that focused on teachers' acquisition of progress monitoring knowledge as well as data management associated with administration, scoring, and decision making for a group of their own students. These measures differed in the algebra skills addressed as well as the format used, which is described in Foegen et al. (2008). Images of the entire first page of each of the three types of algebra measures used in this study are included in [Supplementary materials 1–3](#).

Development of the professional development online system

The online system for the current project was developed using an existing tool (i.e., ThinkSpace¹) to support case-based and critical thinking instruction in higher education (Danielson et al., 2007; Bender and Danielson, 2011; Kruzich, 2013; Wolff et al., 2017). The PD system included two hubs, which are separate features of the system navigable to and from the homepage. One hub comprises the teacher PD; the other hub organizes the data management activities for entering and scoring student data and for tracking progress. Researchers worked with the developer of this platform to adapt the original tool, first creating six asynchronous modules in the PD hub to support teacher learning about algebra progress monitoring and then creating the data management hub with five asynchronous modules to support teachers' management, review, and decision making using graphed progress monitoring data and diagnostic tools.

The content for the first six PD modules mirrored the content for the in-person workshop for practitioners and incorporated multimedia presentations of information (i.e., videos, transcripts) and interactive activities. Within each online module, interactive activities included self-check questions where user answers were followed by expert responses for comparison. All three modules that provided instruction on administration and scoring for each of the three types of algebra progress monitoring measures used a simulated administration of the measures that teachers completed to better understand the student experience. In addition, teachers engaged in scoring exercises in which the modules guided teachers through scoring procedures and provided samples of student work to score as well as an answer key. At the conclusion of each of these modules about the algebra measures, teachers completed a check-out exercise of their scoring accuracy that included automated evaluation of their scoring responses for a completed sample student paper. Teachers were required to achieve at least 90%

accuracy in scoring before continuing with the next module. Additional feedback and practice opportunities were available within the system as well as additional testing opportunities to meet the criterion of 90% accuracy if it was not reached on the first attempt. Following the modules about administration and scoring, we used the same format and approach to develop five new modules to help teachers learn about the online system's data management and decision-making features. Teachers completed all modules asynchronously at times convenient to them. The online PD modules listed the duration of the videos on each page; total module video time ranged from just under 9 to 34.33 mins (for the Algebra Content Analysis module that involved partial credit scoring based on a rubric). Total video time for the 11 modules was 2.59 h; we estimated additional time for teachers to complete activities within each module would add approximately two additional hours for a total time of 4.5–5.0 h.

Prior to the current study, we used an iterative development process across 4 years that included two rounds of "in-house" testing completed by undergraduate preservice teachers or graduate students in mathematics education or special education, followed by testing with four cohorts of teacher participants. Holistic ratings and page-by-page comments were gathered to obtain users' views, and they provided feedback on the content, clarity, visual appeal, and usability of each module. Researchers used this feedback to make refinements to the system and to test the extent to which the system functioned as intended. Although new modules were developed sequentially with several added as each new teacher cohort tried the system, participants evaluated all modules completed to that point in time, so the refinements that researchers made to earlier modules were evaluated by subsequent cohorts. All 11 asynchronous modules were evaluated at least once prior to their use in our final study. The current PD study examined the online PD system by requiring teachers (a) to use all 11 revised modules independently and (b) to administer, score, and use the data management system for at least 10 weeks with at least one class of students taking algebra-related content.

Study design

The current study was funded as a part of a research development grant that required use of a multi-year iterative development process. This process emphasized teacher feedback as the most critical aspect of the project's evaluation, which included feedback about the usability and potential utility of the system. Student progress monitoring is, indeed, an evidence-based practice that teachers may use to inform instructional planning and to effect student gains in achievement, particularly with students who are low achieving (Stecker et al., 2005; Jung et al., 2018; Fuchs et al., 2021); however, teachers who merely collect data and do not do anything differently instructionally

¹ <https://www.thinkspace.org>

based on student data patterns are not likely to effect greater student achievement (Stecker et al., 2005). Consequently, it is important that teachers learn about progress monitoring and how to use data to make meaningful decisions about the adequacy of student progress and the potential need for intervention. Beyond functionality of the system, teacher satisfaction with the PD system, both with instructional features and with the data management tools, remains a necessary first step for its effective use.

Following this iterative process of development, testing, and revisions, the research team conducted a final study of the entire system. District special education directors and curriculum coordinators sent email invitations to general education algebra and special education teachers on behalf of the research team. Interested teachers spoke with research staff for further information and clarification about study components, and the resulting volunteers became the primary research participants. Participating teachers from Iowa, Minnesota, and South Carolina completed 11 instructional online modules on their own. These modules focused on progress monitoring features; three types of algebra progress monitoring measures, including administration and scoring guidelines; and data management, including data entry, graph interpretation, and skills and error analyses. Teachers took a pre- and posttest about their progress monitoring knowledge, provided feedback at three points during their interaction with the modules, and responded to a written questionnaire at the end of the study. Following completion of the online instructional modules and in consultation with researchers, teachers were expected to administer one researcher-assigned algebra measure each week across a period of 10 weeks to at least one class of their algebra students. In addition, teachers administered one self-selected assessment of the two remaining measures to the same students across the 10 weeks and administered the third measure four times, during the first and last 2 weeks of the project. Consequently, the teacher-selected groups of students for whom they administered and scored measures and viewed progress became the secondary research participants. Because the focus of this research was on the teachers' use of the PD system, student participation was necessary for teachers as they considered their students' performance and provided feedback about the system.

Participants

Teachers

A total of 29 teachers participated in this study and completed all training, including 12 teachers in SC, 14 teachers in Iowa, and 3 teachers in Minnesota. Initially, 4 teachers had been recruited in Minnesota, but 2 discontinued their participation in the study shortly after it started, and a third

teacher was recruited through nomination. Of these 29 teachers, 16 were special educators (7 in SC and 9 in IA) and 13 were general educators (5 in SC, 5 in IA, and 3 in MN). See [Table 1](#) for demographic information for each teacher, including the number of years spent teaching, years teaching algebra, type of teacher certification held, gender, and ethnicity.

Students

Students who took the algebra progress monitoring measures ($N = 460$) spanned grade levels from 7 to 12, with the majority of students attending high schools. The types of courses represented included 7th- and 8th-grade General Math, Pre-Algebra, Algebra/Geometry Foundations, Skills and Instructional Strategies, and Algebra 1. Students were typically developing, or they had Individualized Education Plans (IEPs) with goals in mathematics. Students with IEPs received mathematics instruction in inclusive classrooms or received instructional support in algebra by special education teachers in special education settings. The number of students involved, including the number of students with IEPs, and the types of courses in which teachers conducted the progress monitoring activities can be found in [Table 2](#).

Demographic information for providing a general profile for the school or district is provided in [Table 3](#). The number of teachers participating in the research in each school; grades included in the school; student enrollment; and percentages of students with diverse backgrounds, receiving free/reduced lunch (a common proxy for low-income households in the United States), learning English language, and with IEPs are summarized for each school or district according to available data.

Dependent measures

Teacher knowledge and accuracy Knowledge pre- and post-test

For this study, teacher knowledge about progress monitoring and the use of the data-based system was evaluated twice: prior to the start of the instructional modules and again after teachers had completed online instruction, 10 weeks of data collection, and data management. Researchers developed the knowledge test, which was comprised of 25 multiple-choice items with four, possible answer selections (see [Supplementary material 4](#) for the actual assessment used). In addition to assessing general knowledge about progress monitoring, specific items related to administration and scoring of the three algebra measures and the use of the data management system were included. Test items were scored as either correct or incorrect. Cronbach's alpha for the items on the knowledge pretest was 0.86, indicating acceptable internal consistency. Cronbach's alpha for items on the knowledge posttest was 0.84.

TABLE 1 Demographic information for participating teachers.

Teacher	School	Gender	Position ^a	Ethnicity	Years teaching	Years teaching algebra	Teaching certification ^b
1	A	F	GenEd	Caucasian	12	12	SMath
2	B	F	SpEd	Other	13	8	EE, LD, BD, ID
3	C	F	GenEd	Caucasian	23	23	SMath
4	D	F	SpEd	Caucasian	20	20	LD
5	D	F	SpEd	Caucasian	29	20	LD
6	E	F	SpEd	Caucasian	15	5	ECE, EE, LD, BD, ID
7	E	F	GenEd	Caucasian	8	6	SMath
8	E	F	GenEd	Caucasian	2	2	SMath
9	E	F	SpEd	Caucasian	8	4	LD, BD, ID
10	F	F	GenEd	Caucasian	6	2	SMath
11	F	F	SpEd	Caucasian	4	1	EE, LD
12	G	F	SpEd	Caucasian	25	3	LD, ID
13	H	F	SpEd	Caucasian	16	0	EE, LD
14	J	M	SpEd	Caucasian	6	5	EE, SEG
15	I	F	GenEd	Caucasian	-	-	-
16	H	F	SpEd	Caucasian	17	0	SEG
17	I	F	SpEd	Caucasian	6.5	0	ECE, EE, SEG
18	H	F	GenEd	Caucasian	15	15	SMath
19	H	M	SpEd	Caucasian	4.5	0	SEG
20	I	F	SpEd	Caucasian	23	0	LD
21	H	F	GenEd	Caucasian	20	20	SMath
22	K	F	SpEd	Caucasian	8	0	EE, SEG
23	I	F	GenEd	Caucasian	-	-	-
24	I	F	GenEd	Caucasian	23	20	MMath, SMath
25	L	M	SpEd	Hispanic	8	6	SMath, SEG
26	L	F	SpEd	Caucasian	21	13	SMath, BD
27	M	M	GenEd	Caucasian	-	-	-
28	O	M	GenEd	Caucasian	10	10	MMath, SMath
29	N	F	GenEd	Caucasian	6	6	MMath, SMath

^aGenEd = General Education, SpEd = Special Education.

^bBD = SpEd Behavior Disorders, ECE = Early Childhood Education, EE = Elementary Education, ID = SpEd Intellectual Disabilities, LD = SpEd Learning Disabilities, MMath = Middle School Math, SEG = SpEd General, SMath = Secondary Math.

Accuracy of scoring and data entry

Although teachers collected data from student progress measures for 10 weeks as a part of the project, the focus of this study was on teachers' use of the system rather than student performance. To determine the extent to which teachers could learn from the online modules about scoring and data entry, however, researchers included accuracy checks of teacher scoring of their student progress measures as well as accuracy of data entry in the online data management system (See section "Data analysis" for information about procedures for determining scoring and data entry accuracy).

Teacher use and satisfaction

Module feedback

Similar to the earlier iterative cycles of development, instructional modules included feedback pages in the final online PD at several points during the study in which teachers

responded to Likert-type scales and open-ended items. Teachers were asked about the quality of features of the online system, ease of navigation, and their level of engagement during the instruction. They also responded to items about the content of the modules and their level of understanding. In addition, they judged the appropriateness of their time spent in instruction and offered suggestions for revisions that potentially could improve the system or their learning.

Final questionnaire

At the conclusion of the project in their final meeting with a researcher, teachers completed independently a written questionnaire that required holistic ratings and written responses about time they spent looking at student data during the project, tasks in which they engaged across the training and research, any instructional decisions they made based on the data they collected, and specific features about the online system.

TABLE 2 Demographic information for participating students and their courses.

Teacher	School	Course	Students (N)	Students on IEPs (N)	Course description	Grade(s) taught
1	A	Algebra I	29	13	GenEd	9–11
2	B	Tutorial I	11	11	SpEd	9
3	C	Algebra I	22	1	GenEd	9
4	D	Academic Support	9	9	SpEd	10
5	D	Academic Support	8	8	SpEd	9–12
6	E	Academic Support	9	9	SpEd	9
7	E	Algebra I	17	0	GenEd	9–10
8	E	Algebra I-CP	13	0	GenEd	9
9	E	Academic Support	7	6	SpEd	11
10	F	Algebra I AB	15	7	GenEd	9
11	F	Algebra I	16	7	SpEd	9
12	G	Academic Advancement	9	7	SpEd	10
13	H	Skills	4	2	SpEd	9–10
14	J	Pre-Algebra	6	6	SpEd	9–12
15	I	Math 7	22	3	GenEd	7
16	H	Skills	4	4	SpEd	9–12
17	I	Math 7	19	3	SpEd	7
18	H	Algebra 1	17	0	GenEd	9–11
19	H	Skills	4	2	SpEd	9, 10, 12
20	I	Math 8	23	5	GenEd	8
21	H	Algebra 1	13	1	GenEd	9–12
22	K	Resources	11	11	SpEd	7–11
23	I	Math 7	20	1	GenEd	7
24	I	Math 8	12	12	GenEd	8
25	L	Algebra/Geom. Foundations	23	1	At Risk	10–12
26	L	Algebra/Geom. Foundations	47	0	At Risk	10–12
27	M	Math Resources	55	10	At Risk	9–12
28	O	Algebra 1 Lab	12	0	At Risk	9–12
29	N	TransMath 2	47	0	At Risk	8

Procedures

Meetings

Prior to participation in the module training, researchers held an individual face-to-face meeting with each teacher, except the one teacher who was recruited later in Minnesota and met virtually with a project staff member. Following a common outline, researchers presented information about the study, teachers were given a checklist of weekly responsibilities, and they took the knowledge pretest. At the end of the project, staff met again individual teachers who took the knowledge posttest and completed a written questionnaire.

Online professional development

Eleven online instructional modules provided the content for teachers to learn about progress monitoring in general and, more specifically, how to give and score three types of algebra

progress monitoring measures, as well as how to use a custom-designed data management system to record and summarize student graphed data and analyses of their skills and errors. As teachers worked through the online PD on their own, they gave feedback at three points during the modules (early, middle, and endpoint), in which they responded to questions about specific features of the system. They also were able to add comments in each module at any point during their training. Teacher comments and feedback pages were intended to give researchers information about features that seemed to work well as well as any glitches or problems encountered, so any future revisions of the modules could incorporate this information.

Beginning online modules and first evaluation

The first six PD modules focused on content and activities related to progress monitoring concepts and practices and the three algebra measures included in the online system. The first two modules provided the background information for progress

TABLE 3 Demographic information for participating schools or districts.

School (state ^a)	Teachers (n)	Grades served	School size	Student diverse backgrounds (%) ^b	Student free/reduced lunch (%)	Student ELL (%)	Student IEPs (%)
A (SC)	1	9–12	290	5.5	58 (A, B, C districts)	-	15 (A, B, C districts)
B (SC)	1	9–12	960	23		-	
C (SC)	1	9–12	745	10		-	
D (SC)	2	9–12	1025	21	49 (D, E districts)	4 (D, E districts)	12.3 (D, E districts)
E (SC)	4	9–12	700	11			
F (SC)	2	9–12	1760	38	58 (F district)	-	12.5 (F district)
G (SC)	1	9–12	820	24	45 (G district)	-	13 (G district)
H (IA)	5	5–8	420	3	33	0	15
I (IA)	5	9–12	525	3	8	0	10
J (IA)	1	9–12	220	<1	35	<1	12
K (IA)	1	7–12	275	<1	40	0	7
L (IA)	2	9–12	2100	43	63	6	16
M (MN)	1	7–12	400	3.5	36	0	9.5
N (MN)	1	9–12	320	15.5	44	<1	13.5
O (MN)	1	9–12	1050	4.5	21	<1	8

All enrollment numbers and demographic percentages are approximate. School-level data unless marked as district; some school-level or district-level data were unavailable.

^aSC = South Carolina, IA = Iowa, MN = Minnesota.

^bDiverse backgrounds refers to race, culture, and ethnic backgrounds.

monitoring and development research of the algebra measures. The first module, *Core Concepts*, focused on central ideas about the purpose of progress monitoring, its history, and basic features. The *Project AAIMS* module described the development of the progress monitoring tools during a previously funded federal research project. Following completion of these first two modules, teachers completed the first, or early, round of teacher feedback on these two beginning modules.

Middle set of online modules and second evaluation

The next four modules addressed the specific algebra measures. The *Measures Introduction* module provided information common to all three of the algebra progress monitoring measures included in this PD system. The next three modules presented information specific to administration and scoring of each algebra tool: *Algebra Basic Skills*, *Algebra Foundations*, and *Algebra Content Analysis*. In these three modules, teachers had the opportunity to take a measure themselves, so they could experience what would be expected of a student. Teachers learned conventions for scoring each type of measure and had a couple of opportunities to score sample student measures with feedback provided by the system on their accuracy. Teachers had to earn at least 90% accuracy with scoring a type of measure before being allowed to move to the next module. Additional practice and retest options were available. Following their learning and scoring

of these three algebra measures, teachers completed the evaluation feedback page, responding to the same items for this second set of modules.

Last set of modules and third evaluation

The last five online modules focused on features of the data management system. The first of these modules, *Introduction to Data Management*, described the overall capabilities of the system, especially how to add classes or individual students to the database, how to edit student data and make adjustments when students were absent, check student progress, and examine reports that could be generated. The next module, *Evaluating Student Progress*, focused on how to input student scores for the measures and how to view and interpret corresponding student graphs of progress. The *Instructional Decision Making* module showed teachers how to document instructional changes on the graph and how to add or change goals. In addition, recommendations for how to determine the efficacy of the instruction by evaluating graphed student progress were described. Although teachers (rather than the system) scored student performance on the algebra measures, teachers could input total scores as well as item-level data for use in aggregating information about skill proficiency and the common errors students were making. The *Skills Analysis* module showed teachers how data on problem types (i.e., skills) were aggregated for display. Skill reports could be generated to show skill

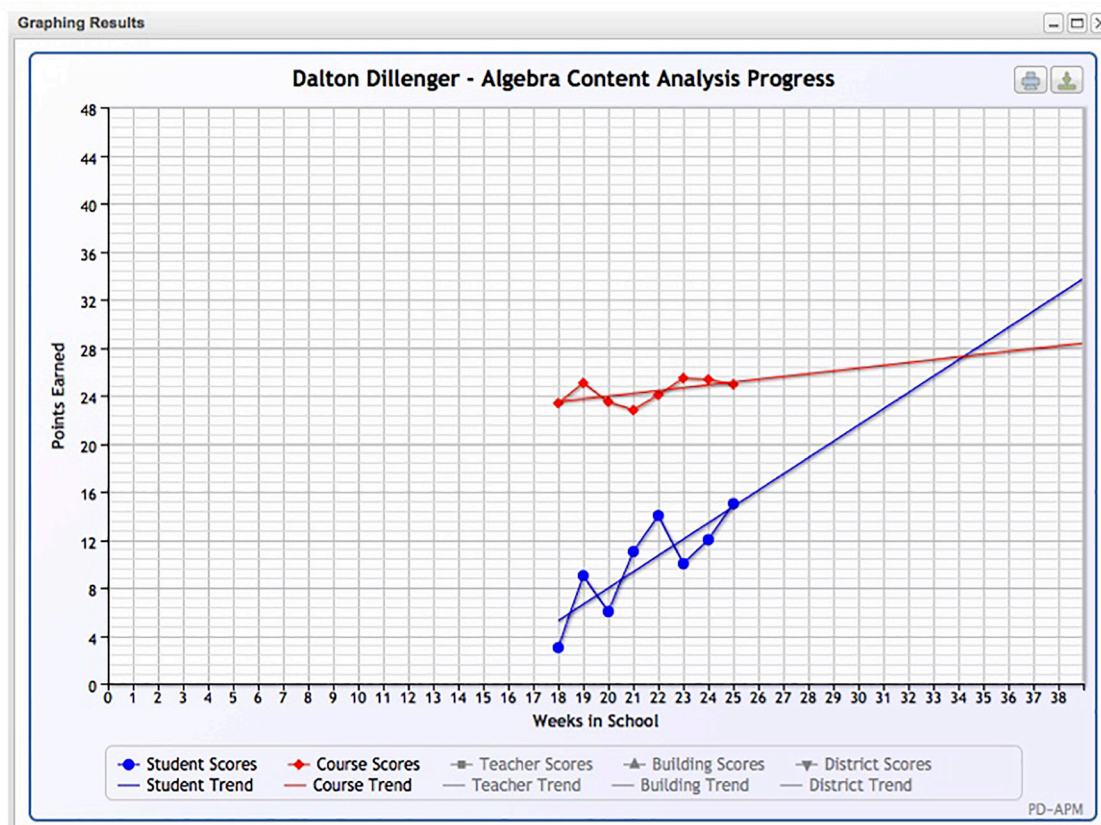


FIGURE 1
Progress monitoring graph showing student data and class comparison data.

proficiency for an individual or for a class of students. The *Error Analysis* module explained how teachers could choose from a list of common errors to note a potential misunderstanding a student made with an incorrect response. Although teachers made the judgments about potential student errors, a drop-down menu of common errors facilitated teachers' data entry. An error analysis report could be generated to depict individual or classwide information, as long as the teacher had entered this item-level data. Finally, teachers completed the last evaluation page for the third set of instructional modules.

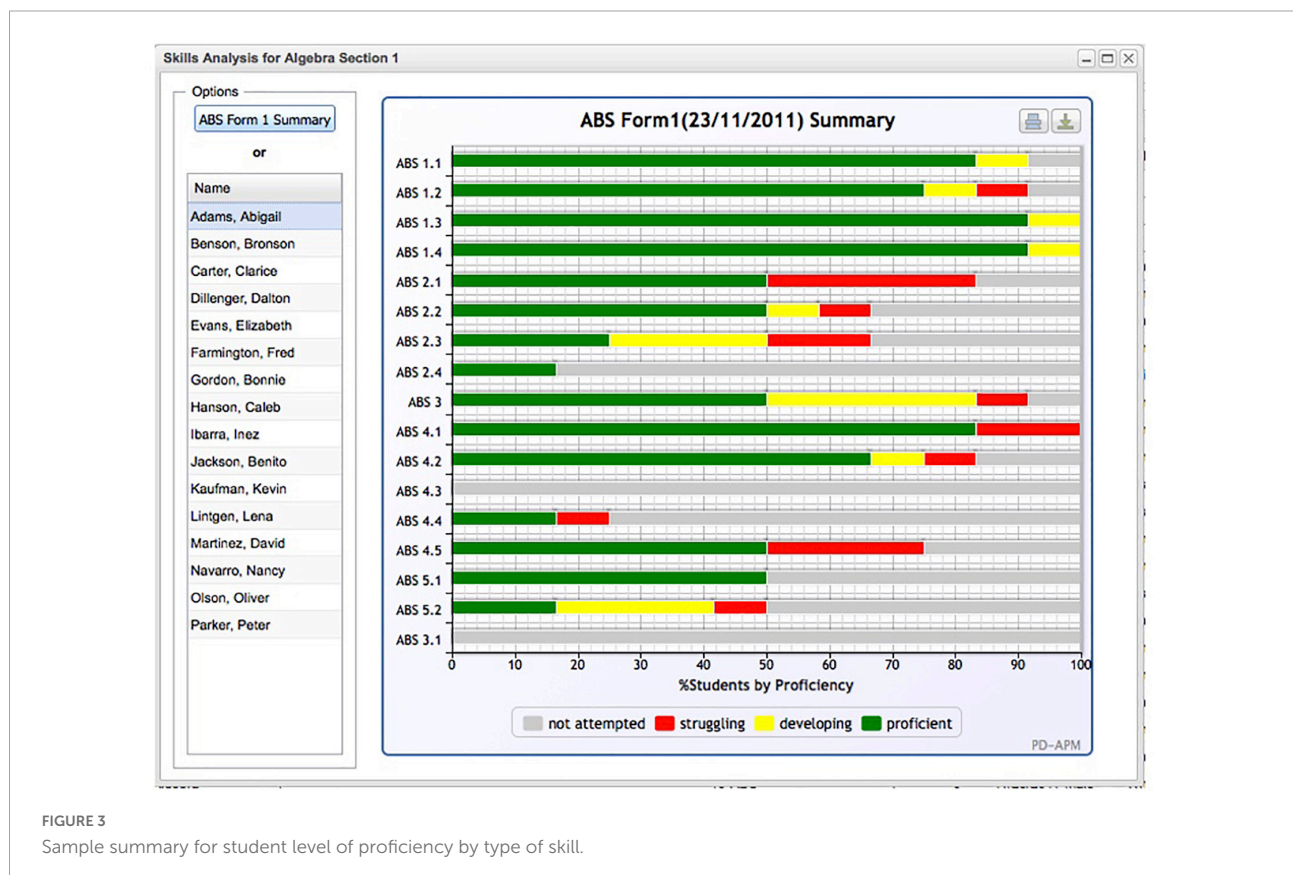
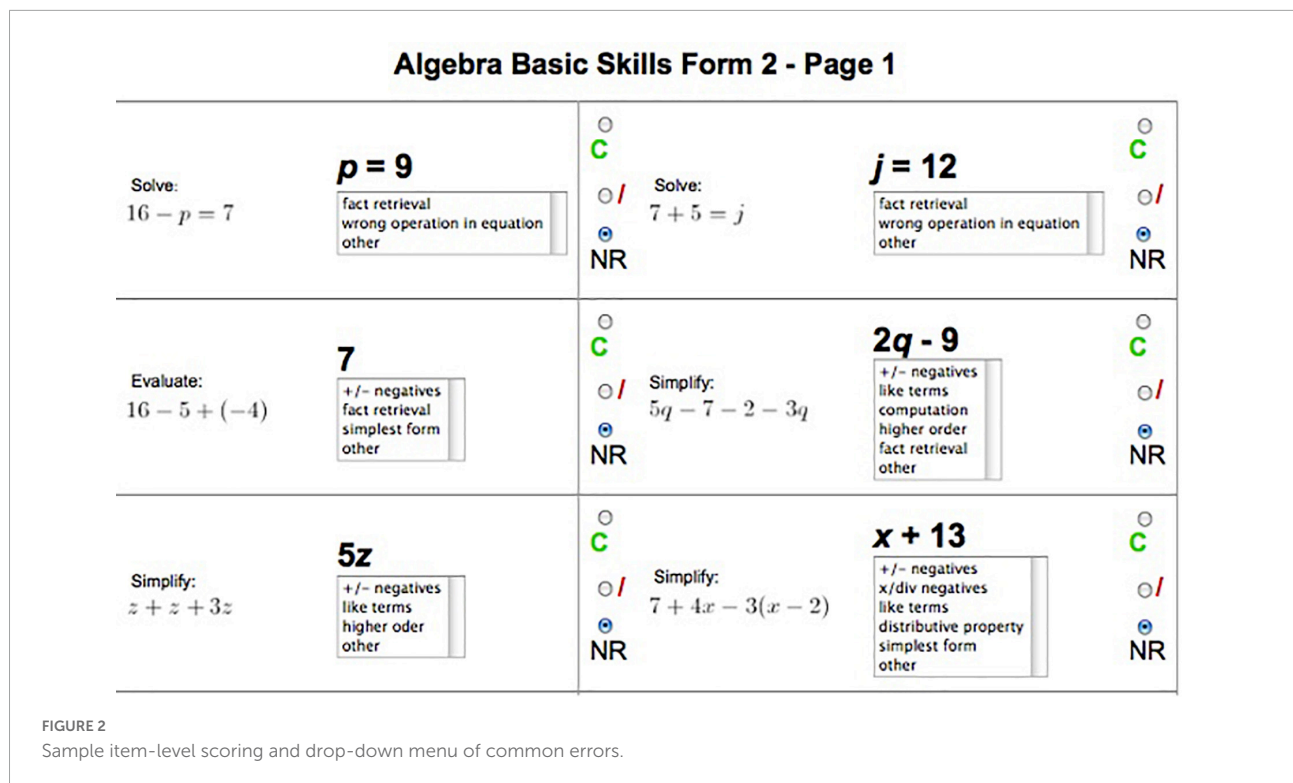
Data management and optional features

Once teachers had completed the modules on their own, they were given access to the Data Management hub and could proceed to add students to course sections, input measures used, and enter student data. After several scores were entered, the data management system could generate a student graph and show the trend of student progress. Teachers had the option to set goals for future achievement and to include phase-change lines to indicate when an instructional modification to the student's program was made. Graphs depicted individual student data points (i.e., scores) and the trend of student

progress but also could show the average score and average trend across the entire class. Figure 1 shows a student's progress monitoring graph with trend line compared to the course average scores and trend.

Progress monitoring

To make sure that all three types of algebra measures were administered and scored during the project, researchers assigned one of the three measures to each teacher, giving consideration to the type of course each teacher selected to monitor and the teacher's preferences. Then teachers were allowed to select a second measure themselves. Teachers gave these two measures weekly across 10 weeks to the entire class. They were required, however, to score student performance and enter data into the data management system for only the primary measure. In addition, teachers administered the third type of progress monitoring measure but only during the first 2 and last 2 weeks of the 10-week period, primarily as a way to document student growth in another way and for researchers to examine relations among the types of measures. Although allowed, teachers were not required to score either their



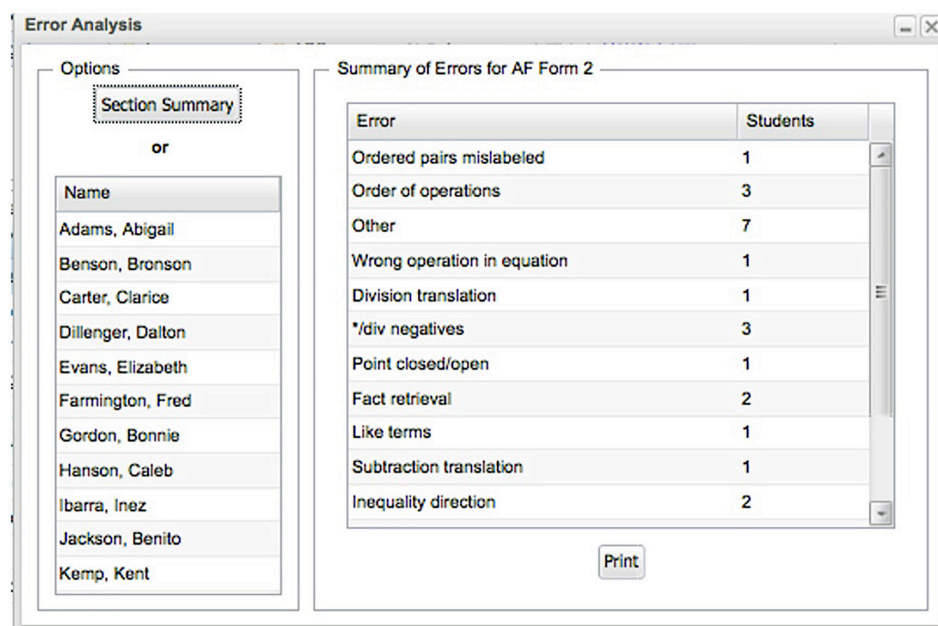


FIGURE 4
Sample report for common errors across class.

secondary or tertiary measures. All measures were turned in to project staff.

Across the teachers, 16 teachers administered Algebra Basic Skills as the primary measure (6 in SC, 8 in IA, 2 in MN), 11 teachers administered Algebra Foundations (4 in SC, 6 in IA, 1 in MN), and 2 teachers administered Algebra Content Analysis (2 in SC). For the secondary measures, 6 teachers administered Algebra Basic Skills, 13 teachers administered Algebra Foundations, and 10 teachers administered Algebra Content Analysis. For the tertiary measure (given during first and last 2 weeks only), 6 teachers administered Algebra Basic Skills, 5 teachers administered Algebra Foundations, and 18 teachers administered Algebra Content Analysis.

Analysis of skills and errors

For required progress monitoring activities, teachers gave the primary assessments weekly to at least one class of students, scored their performance, and entered total correct responses into the data management system. However, for two students in the group, teachers were required also to enter item-level data, that is, accuracy for each student response, and indicate a possible reason for the error for any item answered incorrectly, if they were able to determine one. In this way, teachers had practice using these components of the online system without having to enter data for all items for every student. For this more fine-grained, item-level data entry, teachers were encouraged to select two students who were lower achieving or who had Individualized Education Plans (IEPs). When teachers entered

item-level data (see Figure 2 for sample screen of item-level data entry), the online system was able to generate individual reports (or classroom when applicable) about level of proficiency on each type of skill evaluated on the measure (i.e., proficient, developing, struggling, or not attempted). See Figure 3 for illustration of an individual skills proficiency report. When teachers marked an item as being incorrect, they could choose from a drop-down menu the type of error the student made in that problem, or they could type in an error pattern if they did not see it listed. The system could generate a report of common errors made by student (or by class, if applicable). See Figure 4 for sample common errors report for an individual student.

Data analysis

Teacher knowledge tests and ratings

Teachers' answers on the multiple-choice knowledge pre- and posttests were scored as correct or incorrect, and a matched-pairs *t*-test was used to examine gains. For teacher ratings, descriptive statistics, including frequency counts and/or means, were used to summarize teacher feedback.

Scoring of module practice activities and student measures

To determine whether online instruction was successful in instructing teachers in scoring conventions, we examined the accuracy of their scoring during the interactive practice

activities they did during online instruction about the three algebra measures. At the end of each module that addressed a type of algebra measure, teachers had a check-out exercise for a hypothetical student Max, in which they had to reach at least 90% accuracy in scoring to be allowed to move to the next module.

To check reliability of scoring with the assessments that teachers gave to their students, researchers required teachers to turn in scored papers for the first two test administrations to project staff, who then rescored the entire class. Any disagreements in scoring were discussed with the teacher. Even if the teacher had surpassed the 90% accuracy criterion during the online practice activities, researchers required a 95% accuracy threshold for scoring their own students' measures. Any teacher who did not meet at least 95% for interrater agreement had to return additional sets of their scored measures for an interrater agreement check on all measures until they reached the 95% accuracy threshold. For subsequent administrations after reaching the 95% criterion, researchers rescored a sample of at least 20% of the class measures or a minimum of five assessments for each class administration, whichever was more. When accuracy fell below the 95% threshold, the entire class set of papers was rescored. A few teachers chose to score performance on the secondary and tertiary measures themselves. When they did, their scoring reliability was checked in the same way.

For Algebra Basic Skills and Algebra Foundations, responses were scores simply as correct or incorrect. Consistent with other progress monitoring research (e.g., Fuchs et al., 1994), interrater agreement was calculated as the total number of agreements in scoring divided by the sum of the total agreements and total disagreements. For Algebra Content Analysis, however, students could show work and be awarded partial credit for each of the 16 problems. Interrater agreement was calculated by subtracting the number of scoring disagreements from 16 and then dividing that difference by 16.

Data entry of total scores on primary measures

To determine reliability of teachers' data entry, researchers compared the student scores teachers had recorded on the student measures with the scores they had entered into the data management system. Even if researchers had determined that the teacher had scored a student measure inaccurately and had adjusted that student score for analyses of student data, researchers still compared what the teachers had written directly on the student measures with what the data they entered in the online system. For each class, researchers figured the number of matches between the recorded scores on student papers and the scores entered into the system. The number of matches was divided by the total number of students to determine the interrater data entry percentage of agreement.

Results

Researchers analyzed data to examine the extent to which the online system worked as intended. We examined whether the online system led to improved teacher knowledge and skills with algebra progress monitoring. Researchers evaluated teachers' knowledge through a pre-and posttest. Their accuracy in scoring and data entry were evaluated. Efficiency of the system and teacher satisfaction with instructional modules were examined through teacher self-report information and rating scales. A total of 29 teachers completed the training from beginning to end, administering algebra measures, scoring student performance, entering data in the online management system, and giving feedback. Note that some data were not accessible due to technical glitches with the system or because a few teachers chose not to respond to particular questions.

Teacher knowledge and accuracy

Knowledge test

The same knowledge assessment was given to teachers as a pre- and posttest. Cronbach's alphas for the pretest and posttest were 0.86 and 0.84, respectively, indicating adequate internal consistency. The posttest was administered during the final, wrap-up meeting with project staff. A paired *t*-test indicated that teachers' accuracy improved significantly from pre- to posttest, $t(28) = -7.59$, $p < 0.001$. Means with standard deviations in parentheses for item accuracy on the pretest and posttest were 9.97 (5.02) and 17.66 (2.83), respectively, for this 25-item assessment.

Accuracy in online scoring activities

Teachers had to reach a criterion level of accuracy in scoring the student exercise(s) before moving forward with another module (see Table 4). However, researchers also were

TABLE 4 Scoring accuracy during module exercises for three types of algebra measures.

Module scoring exercise	<i>n</i>	Min. (%)	Max. (%)	<i>M</i> (%)	<i>SD</i>
Algebra basic skills					
Attempt one (max)	27	91.7	100	98.88	2.20
Attempt two (rachel)	5	90.2	100	97.56	4.24
Algebra foundations					
Attempt one (max)	26	92.5	100	97.69	2.44
Attempt two (rachel)	2	97.3	97.3	97.3	0.00
Algebra content analysis					
Attempt one (max)	26	50	100	85.84	10.40
Attempt two (rachel)	13	62.5	100	88.48	10.48

TABLE 5 Interrater agreement for scoring of measures and online data entry.

	Week									
	1	2	3	4	5	6	7	8	9	10
Primary measures	Percent of scoring agreement (n)									
Algebra basic skills	99.0 (15)	98.0 (16)	98.5(17)	98.4 (17)	98.8 (17)	98.9 (17)	98.1 (17)	99.0 (17)	98.9 (16)	99.3 (16)
Algebra foundations	90.3 (8)	95.9 (9)	96.1 (9)	96.2 (9)	97.0 (9)	96.1 (9)	97.0 (9)	97.8 (8)	96.6 (8)	97.6 (8)
Algebra content analysis	96.0(2)	97.0(2)	99.0(2)	96.5(2)	98.0(2)	96.0(2)	95.5(2)	97.5(2)	97.5(2)	98.5(2)
	Percent of data entry agreement (n)									
Algebra basic skills	96.9 (13)	94.2 (16)	94.3 (16)	96.8 (16)	98.6 (16)	95.3 (16)	99.1 (16)	96.5 (16)	97.5 (15)	97.0 (15)
Algebra foundations	90.4 (9)	90.2 (9)	91.9 (9)	93.6 (9)	100 (9)	89.7 (9)	98.9 (9)	95.5 (8)	99.0 (8)	92.3 (8)
Algebra content analysis	96.0 (2)	100 (2)	100 (2)	96.0 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)

TABLE 6 Teacher evaluation of their level of understanding of module content.

Module	Frequency of response by rating					Total responses (n)	M	SD
	1	2	3	4	5			
Early (after Module 2)	1	1	2	11	9	24	4.08	0.61
Middle (after Module 6)	0	1	1	11	7	20	4.20	0.74
End (after Module 11)	0	1	4	11	12	28	4.21	0.84

For the teacher responses, 1 = lowest level of understanding, 5 = highest level of understanding.

interested in the accuracy with which they scored their own student papers. Therefore, project staff evaluated interrater agreement for teachers' scoring on their primary measures. In addition to the scoring accuracy of algebra measures, researchers checked teachers' accuracy for data entry based on teachers' markings of the measures themselves. Table 5 shows percentage of interrater agreement for scoring each of the primary measures across 10 weeks of weekly data collection and the number of teachers engaged each week with those tasks as well as accuracy of their data entry in the online data management system.

Teachers' use of the online system

Ratings of the instructional modules

At three points during the online training (i.e., early, middle, and end), teachers completed the same set of Likert-scale ratings to indicate their level of understanding of the online instructional content on a scale of 1–5, with 1 indicating the lowest level of understanding and 5 indicating thorough understanding. The early evaluation followed the first two modules that focused on critical concepts of progress monitoring and the background research for the development of the three algebra measures to be taught. The middle evaluation took place after the next four modules. These modules introduced the three

algebra measures and then focused on each individually, requiring practice in how to administer and score each type of assessment. The last set of module evaluation ratings took place after the next set of five modules that focused on features of the data management system and data entry of scoring, skill performance, and common errors. Frequencies for the teachers' ratings are found in Table 6.

Efficiency of online modules, administration, and scoring tasks

Teachers were asked an open-ended question about whether they thought the time they spent viewing the instructional modules was reasonable. Researchers also asked for explanations to support their responses. Responses were classified and coded as a "0" if the teacher responded negatively, a "1" if indicating the time was "okay," "somewhat" or another variation indicating moderate satisfaction, and a "2" if responding "yes." Table 7 provides this information across teachers at the three evaluation checkpoints (i.e., early, middle, and end).

Teachers also were asked during the final meeting with researchers to complete a questionnaire containing items about their acceptability with the amount of time they spent in various activities. This Likert-type scale ranged from 1 to 4, with 1 = *completely agree* to 4 = *completely disagree*. Table 8 provides acceptability of time involved in the completion of the

TABLE 7 Teacher responses for “was the time spent on the modules reasonable?”

Module	Frequency of response (<i>n</i>)			Teachers responding
	0	1	2	
Early (after Module 2)	0	3	21	24
Middle (after Module 6)	0	3	17	20
End (after Module 11)	6	8	14	28

For the teacher responses, 0 = no, 1 = somewhat/okay, 2 = yes.

instructional modules, administration of measures, and scoring of measures.

Teacher overall satisfaction with online modules

At three checkpoints during the online instruction, teachers rated their level of satisfaction (1 = low satisfaction, 5 = high satisfaction) with the modules, appropriateness of the modules' level of difficulty, and the teachers' level of task engagement during the modular instruction. Table 9 presents these teacher satisfaction data.

Teachers also rated their level of satisfaction (1 = *low satisfaction*, 5 = *high satisfaction*) with features imbedded in the online PD, such as the quality of graphics in the modules, clarity of module content, organization of the module, and ease of navigation. Table 10 displays the number of teachers who rated each feature by their level of satisfaction with system features.

Additionally, on the final questionnaire, teachers indicated whether they thought the content reflected on the progress monitoring measures was appropriate for their classes. Teachers used a Likert-type scale (1 = *completely disagree*, 2 = *disagree*, 3 = *agree*, 4 = *completely agree*) to reflect their level of agreement: 1 = 1 teacher, 2 = 2 teachers, 3 = 11 teachers, and 4 = 15 teachers, with $M = 3.38$ and $SD = 0.78$.

Use of optional online features

Several features in the data management system were covered in the online PD but were not required for use during the project, such as reviewing student graphs, comparing individual and class progress graphs, examining individual

or class skills information, and examining individual or class common errors. However, some teachers chose to use these optional features during the project. At the final meeting, teachers indicated whether they had used specific system features. Table 11 provides the number of teachers using each data-based decision-making feature that was available but not required to be used during the project period.

Discussion

Teacher knowledge and accuracy

One goal of the study was to determine whether knowledge about algebra progress monitoring could be improved among teachers using the professional development online system and to verify that they could be highly accurate in scoring algebra measures based on the online instruction. Without a comparison group, increases in teacher knowledge must be interpreted cautiously. However, based on the study information, teachers improved significantly on the knowledge assessment about progress monitoring and the use of the online system from pre- to posttest. Teachers grew by an average of almost eight items by posttest. However, actual growth may have been a little greater. At pretest, two of the teachers took the assessment outside of research staff meetings due to complications that arose with scheduling and the distance required for travel and exhibited the highest pretest scores across the entire teacher sample (i.e., scores of 17 and 18). Consequently, the fidelity of these results is unclear.

To determine whether teachers could learn to apply scoring conventions accurately with the algebra measures used, researchers evaluated teacher learning during the practice exercises in the modules (see Table 4). Results from the practice exercises in the PD modules indicated that teachers were successful in learning scoring conventions and applying them to completed student problems. Accuracy for Algebra Basic Skills and Algebra Foundations measures was very high at 99% and 98%, respectively. The Algebra Content Analysis measure, though, required more complex scoring with potential awarding of partial credit for problems exhibiting student

TABLE 8 Final questionnaire: Acceptability of time for professional development (PD), administration, and scoring.

Questionnaire item	Frequency of response by rating				Total responses (<i>n</i>)	<i>M</i>	<i>SD</i>
	1	2	3	4			
“The amount of time I spent completing the professional development modules for this project was acceptable.”	0	2	15	12	29	3.34	0.61
“The time it took to administer the measures to my students was acceptable.”	1	1	9	18	29	3.52	0.74
“The time it took to score the measures was acceptable.”	1	3	7	18	29	3.43	0.83

For the teacher ratings, 1 = completely disagree, 2 = disagree, 3 = agree, 4 = completely agree.

work. Consequently, teachers' accuracy was not as high (i.e., 86%). More teachers completed a second scoring exercise in the module for Algebra Content Analysis measures than they had for Algebra Basic Skills and Algebra Foundations. They improved modestly with this second attempt, but not every teacher achieved the 90% criterion for moving to the next module. When that occasion occurred, researchers met with teachers individually to review scoring procedures, answer questions, and provide support.

Importantly, teachers were highly accurate in scoring their own students' papers (see Table 5). The lowest interrater agreement percentages across all three measures occurred during the first couple of weeks of test administration, indicating that teachers improved their accuracy with additional practice. Although interrater agreement was very high for the more difficult Algebra Content Analysis measure when scoring their own students' papers, only two of the teachers were required to score the Algebra Content Analysis as their primary assessment. Consequently, evaluation of additional teachers scoring Algebra Content Analysis measures is recommended. In addition to scoring student measures, teachers had to enter scores in the online data management system. Teachers were accurate in transferring scores from their student measures to the online system.

Teacher satisfaction and use of the online professional development system

Instructional modules

Researchers asked teachers to rate their level of understanding of the module content at three occasions,

once after the first two modules, after the next four modules, and after the last five modules. Teachers used a Likert-type scale with 1 indicating the lowest level and 5 indicating the highest level of understanding. Mean scores for all three occasions were greater than 4.0, indicating that teachers thought they understood the information being presented. The lowest mean rating (i.e., 4.08) was for the earliest feedback occasion in which the modules being considered included background information about progress monitoring and the research endeavors to support development of the algebra measures. The rest of the modules focused more directly on hands-on tasks for teachers (i.e., giving and scoring the algebra measures and using the data management system) and were rated more highly in terms of their level of understanding.

Several other questions probed teacher satisfaction with the online PD system. At these same three feedback intervals, teachers rated their overall satisfaction with the modules, the appropriateness of the level of difficulty of the modules, and their level of engagement while working through the modules. Likert-type ratings from 1 to 5 were used with "1" indicating the lowest and "5" as the highest satisfaction, appropriateness of difficulty, or level of engagement. Mean scores ranged from 3.88 to 4.40, indicating overall high teacher ratings of the PD modules. With respect to the item about overall satisfaction with the modules, the lowest mean rating (i.e., 3.93) was given for the modules describing the components of the data management system. Corroborating ratings for their level of understanding of the instructional modules, the lowest mean rating for both the appropriateness of difficulty of the instructional modules and teachers' level of engagement during the PD was given for the early modules on the background of progress monitoring and

TABLE 9 Teacher ratings of modules: Satisfaction, difficulty, and engagement.

Item and point in time	Frequency of response by rating						<i>M</i>	<i>SD</i>
	1	2	3	4	5	Total responses (<i>n</i>)		
Item 1: Your overall level of satisfaction with these modules								
Early (after Module 2)	0	0	4	11	9	24	4.21	0.72
Middle (after Module 6)	0	0	2	12	6	20	4.20	0.62
End (after Module 11)	0	0	2	12	6	20	3.93	0.94
Item 2: The appropriateness of these modules' levels of difficulty								
Early (after Module 2)	1	0	3	12	8	24	4.08	0.93
Middle (after Module 6)	0	0	1	10	9	20	4.40	0.60
End (after Module 11)	0	0	4	10	14	28	4.36	0.73
Item 3: Your level of engagement while working on these modules								
Early (after Module 2)	1	1	5	10	7	24	3.88	1.04
Middle (after Module 6)	0	0	4	9	7	20	4.15	0.75
End (after Module 11)	0	1	2	17	8	28	4.14	0.71

For the teacher ratings, 1 = lowest satisfaction/appropriateness/level of engagement, 5 = highest.

research development of the algebra measures. Across these results, the research team inferred that the background and research information may have been a little less engaging and perhaps harder to understand than the other modules focused on information that teachers would use directly with their students or within the data management features. Interestingly, though, teachers reported an overall high level of satisfaction with this same group of modules (i.e., section “Instructional modules”).

At these same three feedback intervals, researchers also asked about the system’s technical features of PD. Teachers were asked to rate from 1 to 5 (i.e., low to high) about organization of the modules, clarity of content, quality of graphics used, quality of animation used, quality of narration, and the ease of navigation through the system. Teachers’ mean ratings were high across all these features for each of the three sets of modules. In fact, mean ratings were 4.0 or higher at each feedback interval for each of the items except one. The item

TABLE 10 Teacher ratings of modules: Organization, navigation, and quality.

Item and point in time	Frequency of response by rating					Total responses (<i>n</i>)	<i>M</i>	<i>SD</i>
	1	2	3	4	5			
Item 1: The organization of these modules								
Early (after Module 2)	1	0	2	6	15	24	4.42	0.97
Middle (after Module 6)	0	0	2	8	10	20	4.40	0.68
End (after Module 11)	0	2	4	8	14	28	4.21	0.96
Item 2: The clarity of the content in these modules								
Early (after Module 2)	0	0	2	11	11	24	4.38	0.65
Middle (after Module 6)	0	1	5	8	6	20	3.95	0.87
End (after Module 11)	0	0	6	9	13	28	4.25	0.80
Item 3: The quality of the graphics used in these modules (clarity, contributes to understanding)								
Early (after Module 2)	0	1	4	5	14	24	4.33	0.92
Middle (after Module 6)	0	1	3	7	9	20	4.20	0.89
End (after Module 11)	0	1	2	9	16	28	4.43	0.79
Item 4: The quality of the animation used in these modules (clarity, audibility, contributes to understanding)								
Early (after Module 2)	0	1	5	8	10	24	4.13	0.90
Middle (after Module 6)	0	2	5	4	9	20	4.00	1.08
End (after Module 11)	0	1	3	12	11	28	4.21	0.79
Item 5: The quality of the narration used in these modules (clarity, audibility, contributes to understanding)								
Early (after Module 2)	0	0	4	8	12	24	4.33	0.76
Middle (after Module 6)	0	0	4	8	8	20	4.20	0.77
End (after Module 11)	1	0	1	12	14	28	4.36	0.87
Item 5: The ease with which you could navigate through the system								
Early (after Module 2)	1	1	3	6	13	242	4.212	1.10
Middle (after Module 6)	1	2	1	7	9	20	4.05	1.19
End (after Module 11)	1	3	0	9	12	25	4.12	1.16

For the teacher ratings, 1 = lowest, 5 = highest.

TABLE 11 Teacher reports of task engagement with data management features.

Questionnaire item	Frequency of response (N = 29)
	Yes
Examining student progress graphs	21
Comparing student and class progress graphs	16
Reviewing progress graphs with students	11
Inserting phase changes	3
Examining individual student skills information	25
Examining class skills information	15
Reviewing skills information with students	13
Examining individual student errors information	27
Examining class student errors information	16
Reviewing errors information with students	18

indicating clarity of content was 3.95 for the middle group of modules that focused on administration and scoring of the measures. In fact, the lowest mean rating (although still high at 4.0 or higher) occurred for this same group of instructional modules regarding the quality of the graphics, animation, and narration used as well as the ease of navigation through the system. This group of modules addressed three different types of algebra progress monitoring measures and taught teachers how to administer and score them. The modules were highly interactive and expected teachers to engage in practice activities. In fact, this set of modules was the only set that required teachers to reach a specified criterion with scoring before proceeding to the next module. It could be that narration, clarity, graphics, and animation were even more critical with these modules, as teachers observed models of the tasks they were to perform. Another possible explanation is that this group of modules included the scoring of the Algebra Content Analysis measures. Based on accuracy data and attempted practice exercises, this measure was harder for the teachers to learn to score successfully. It was the last module in this group of modules prior to completing the feedback, so the recency of this more difficult task may have affected teacher ratings across the entire group of modules.

Efficiency

Another aspect of teacher acceptability for the final version of the online PD system was teachers' judgments of PD efficiency. That is, at the requested three feedback intervals, teachers indicated whether the time they spent working through the online PD was reasonable to them. They responded through an open-ended format, so they could provide context for their

responses. All of the teachers indicated that the amount of time was somewhat reasonable or reasonable at the first two feedback occasions. However, 6 of 28 teachers judged the time spent on the last five modules that focused on data management aspect of the system to be unreasonable. Several teachers reported having internet connectivity problems or being busier with other tasks at this time. Because it was the last set of modules, some teachers explained that they thought the PD could have fewer modules or perhaps more condensed versions of the modules. Also, it should be noted that this last group of modules all focused on the use of the data management system. The research team first converted previous face-to-face PD to the six beginning online instructional modules. The research team then developed the data management system as a part of the overall grant project and created corresponding PD instructional modules to match the new data management system. The last several modules had been viewed by only two teachers prior to the current study. Consequently, the team had not received as much feedback for refinement with these modules as they had received with earlier modules. Additionally, other research corroborates that data-based decision making is difficult for teachers (and preservice teachers) to apply and that their interpretations often are qualitatively different (e.g., less cohesive) from expert users and trainers of progress monitoring (Espin et al., 2017, 2021a; Wagner et al., 2017). More attention may need to be directed to crafting these data-based decision-making modules to make them more explicit and acceptable to teachers. Additional feedback and knowledge checks should be solicited for each of these modules.

Researchers also asked teachers about efficiency on the final questionnaire at the end of the study. Teachers were asked to think back across all the PD as well as the administration and scoring of the progress monitoring measures. On this questionnaire, teachers used a rating scale to indicate their level of agreement with statements about the acceptability of the time they spent completing modules, administering progress measures, and scoring measures. All statements were written in the affirmative (i.e., time in tasks was acceptable), but the ratings forced a choice between agreement and disagreement. The scale was "1" for *completely disagree*, "2" for *disagree*, "3" for *agree*, and "4" for *completely agree*. All 29 teachers responded to these items. Only two teachers disagreed with the statement about the amount of time spent in the instructional modules as acceptable. Thus, across all the teachers and considering the totality of the project, teachers rated their time in the online PD as acceptable.

Of the three ratings about the acceptability of time that it took to complete tasks, teachers rated administration of the student measures with the highest mean rating of acceptability (3.52 of 4.0). However, two teachers either disagreed or completely disagreed with this statement that the time spent was acceptable. These progress measures took either 5 or 7 mins once per week, depending on the particular type of measure given. However, it is possible that these two teachers were

considering that administration of all three measures (when only one was required to be scored and entered into the system) was not acceptable. The majority of the teachers, however, agreed or completely agreed that time spent administering the algebra progress measures and scoring the measures was acceptable, 27 of 29 and 25 of 29, respectively.

Optional features of the data management system

On the final questionnaire at the end of the study, teachers also responded to items indicating whether they had used components of the data management system on their own. That is, using these features was not required as a part of study participation, but the online instructional modules provided information about how to access and use these features. The majority of teachers reported examining student progress graphs (21 of 29), but far fewer actually reviewed the graphs with their students (only 11 of 29). Although teachers had been asked to enter item-level information from the measures for only two of their lower performing students, some teachers chose to enter skills and/or common errors information for more of their students or their entire class. In fact, almost all of the teachers (27 of 29) reported examining individual student errors information, with 18 teachers reviewing common error information with their students, and 16 teachers examining student errors across their class. Similarly, 25 of 29 teachers examined individual skill proficiency (i.e., level of mastery for skills included on the measures), with 13 teachers reviewing the skills information with their students, and 15 examining skills information across their class. The activity in which the fewest teachers engaged was inserting phase change lines on student graphs. When asking about time spent viewing student data each week, teachers reported a range of 5–150 mins with a mean of 45 mins. Thus, teachers appeared to take advantage of the available data management tools in the online system even when not required to do so.

Summary of results and future research

Conclusion

With this online PD system, teachers acquired knowledge and skills about how to conduct progress monitoring in algebra. They scored student algebra progress measures accurately and entered data successfully into the online management system. Teachers reported overall high levels of satisfaction with the modular training, including the content, difficulty level, and organization of the instruction as well as the clarity of the imbedded technological features. They were able to access the system's data management components and store student data. They reported that the time spent in the PD activities, including the instructional modules and the administration and scoring of student measures, was acceptable to them.

Overall, the development and implementation of an online PD system for instructing teachers in how to conduct and manage algebra progress monitoring appeared successful. It functioned as intended and enabled 29 general education and special education teachers to learn to give and score three types of algebra progress measures as well as store and view student data across time.

Study limitations

A number of limitations should be noted with this study. First, the study required teachers to report their satisfaction with the system, which could be positively biased. Second, a pretest/posttest design was used to collect information. Without a comparison group, it is difficult to judge fully the efficacy of the PD. Third, not all teachers responded to all requests for feedback. Although teachers viewed an online evaluation page at three points during the online instruction, teachers could proceed to the next module even if they failed to complete some (or all) of the items. Additionally, occasional internet connectivity issues at schools or teachers' homes sometimes made access difficult or interfered with particular tasks. Fourth, although researchers were able to calculate accuracy for the knowledge test and scoring of student measures, direct observations of teachers working through online modules, administering measures in the classrooms, or using the data management system were not conducted. Of course, the overall purpose of the study was to determine whether teachers could learn to conduct algebra progress monitoring on their own. However, teacher self-report responses, with unknown reliability, provided the majority of the data for this project evaluation. Fifth, teachers were not required to use all the available components of the data management system in this evaluation study. Consequently, researchers received only anecdotal information about some of the available online features. Sixth, although some schools had multiple teachers using the online system, we did not evaluate systematically whether teachers completed all activities independently or whether they discussed features with one another. Last, researchers experienced a several-month delay in getting all features of the online system fully functional. Although all teachers completed all the online PD modules, depending on how quickly they worked through modules independently, some teachers had to begin administration of algebra progress measures prior to completing the modules related to data management. That is, they needed to administer progress measures for 10 weeks, and several teachers would have run out of time in the school year if they had waited until completing all modules before administering the 10 weeks of assessments. Therefore, some teachers did not enter data into the management system on a weekly basis; instead, they grouped batches of assessments (especially the first few weeks of data) to enter at one time after they had completed the modules about using the data management system, which may have affected

their reliability in scoring. Relatedly, during the final meeting in which teachers completed a questionnaire about their overall satisfaction with the PD, some teachers reported anecdotally that it had been a long time since they had worked through the modules, while others said they had completed all modules closer in time to the final meeting. It is not known how this variation in length of time spanned to complete all the modules may have affected teachers' responses about the PD modules and the related assessment activities or how the time they had left after completing data management modules affected their interest in exploring the data management features that had not been required to be used during the project.

Implications for future work

Although teachers were asked to enter problem-by-problem accuracy on measures for two students in their classes and indicate a possible error when the student's response was inaccurate, some teachers chose to enter item-level data for their entire class. At first glance, the assumption could be made that teachers understood the potential benefits of such a data management system for ongoing progress monitoring. However, fewer teachers reported viewing individual student graphs of total scores, and less than half the teachers reported showing graphs to students. In fact, more teachers reported examining student skills and errors in the data management system and showing these graphics to students than examining and showing student graphs of progress monitoring scores of measures across time. It may be that teachers recognized how knowing about proficiency of algebra skills and the common errors students made could assist them as they decided how to alter instruction for their students. However, the basic tenets of progress monitoring that include decision making about instructional effectiveness tied to judgments about student rate of improvement may not have been realized by all the teachers or perhaps not emphasized enough in the PD. With progress monitoring, technically sound data should be used for ongoing instructional decision making, especially for determining when student progress is not adequate for meeting goal expectations. An equally important aspect in data-based individualization is the use of available progress monitoring data and other diagnostic data to determine the nature of the instructional modifications to better meet individual needs. Consequently, implementation of this PD system may need to include more specific content about both instructional decision making and appropriate intensification of intervention, especially for individual students who are not progressing as expected.

Future research with this online PD should include systematic evaluation of all the data management components of the system. Teacher evaluation of each module could be required prior to navigation to subsequent modules. Features that were optional for teachers or minimally required in

the current study should be evaluated further. Recognizing that teachers frequently need support to make the best use of progress monitoring data and instructional decision making (Stecker et al., 2005; Espin et al., 2017; Wagner et al., 2017; Jung et al., 2018; Fuchs et al., 2021), a module that includes additional information focused on data interpretation and instructional utility may need to be developed. Perhaps a module for administrators or lead teachers could assist school staff if implementation of procedures were adopted for particular courses. Exploring how in-person or online data team meetings might be used effectively to support teacher decision making is another aspect that could be examined. In addition to refining data utilization aspects, the current online PD could include support for teachers about generally effective algebra instruction and how to intensify instruction when students continue to struggle.

The PD system also could be adapted easily for use with other areas of readily available mathematics measures, such as those for elementary and middle school levels in computational fluency and concepts/applications or problem solving or for early numeracy (e.g., number identification, quantity discrimination, missing number). It could be expanded for use with progress monitoring in other academic areas, such as reading, writing, and discipline-specific vocabulary.

Next steps for online professional development

The development of the online PD and data management system was led by the faculty member who originally developed ThinkSpace (Bender and Danielson, 2011). Following his retirement, a small company took over the development leading to the version used in this paper and in a subsequent research project. Due to transitions within the company, along with transitions at the university level, efforts to shift the system from the cloud system used by the developer to the university's information technology system have required more time than was anticipated. New opportunities for completing this process have become available, and we anticipate that this online PD system will be moving toward wider accessibility in the near future.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Iowa State University Institutional Review

Board; Clemson University Institutional Review Board. The teacher/participants provided their written informed consent to participate in this study and aggregate information about students in their classes.

Author contributions

Both authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

Funding

The research reported here was supported by the Institute of Education Sciences, United States Department of Education, through Grant R324A090295 to Iowa State University as part of the Professional Development for Algebra Progress Monitoring Project.

Acknowledgments

The authors wish to acknowledge Jeannette Olson, Vince Genareo, Amber Simpson, and Renee Lyons for their contributions to the development and evaluation of the online PD system.

References

- Bender, H. S., and Danielson, J. A. (2011). A novel educational tool for teaching diagnostic reasoning and laboratory data interpretation to veterinary (and medical) students. *Clin. Lab. Med.* 31, 201–215. doi: 10.1016/j.cl.2010.10.007
- Danielson, J. A., Mills, E. M., Vermeer, P. J., Preast, V. A., Young, K. M., Christopher, M. M., et al. (2007). Characteristics of a cognitive tool that helps students learn diagnostic problem solving. *Educ. Tech. Res. Dev.* 55, 499–520. doi: 10.1007/s11423-006-9003-8
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Except. Child.* 52, 219–232. doi: 10.1177/001440298505200303
- Espin, C., Chung, S., Foegen, A., and Campbell, H. (2018). “Curriculum-based measurement for secondary-school students,” in *Handbook on Response to Intervention and Multi-Tiered Instruction*, eds P. Pullen and M. Kennedy (New York, NY: Routledge), 291–315.
- Espin, C. A., Förster, N., and Mol, S. E. (2021a). International perspectives on understanding and improving teachers’ data-based instruction and decision making: Introduction to the special series. *J. Learn. Disabil.* 54, 239–242. doi: 10.1177/00222194211017531
- Espin, C. A., van den Bosch, R. M., van der Liende, M., Rippe, R. C. A., Beutick, M., Langa, A., et al. (2021b). A systematic review of CBM professional development materials: Are teachers receiving sufficient instruction in data-based decision making? *J. Learn. Disabil.* 54, 256–268. doi: 10.1177/0022219421997103
- Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., and de Rooij, M. (2017). Data-based decision-making: Developing a method for capturing teachers’ understanding of CBM graphs. (2017). *Learn. Disabil. Res. Pract.* 32, 8–21. doi: 10.1111/ldrp.12123
- Foegen, A., Olson, J., Genareo, V., Dougherty, B., Froelich, A., Zhang, M., et al. (2017). *Algebra Screening and Progress Monitoring data: 2013-2014 (Technical Report 3)*. Ames, IA: Iowa State University.
- Foegen, A., Olson, J., and Impeccoven-Lind, L. (2008). Developing progress monitoring measures for secondary mathematics: An illustration in algebra. *Assess. Effect. Interv.* 33, 240–249. doi: 10.1177/1534508407313489
- Foegen, A. (2004-2007). *Project AAIMS: Algebra assessment and instruction—Meeting standard (Award # HC324C030060)*. [Grant]. Washington, DC: U. S. Department of Education.
- Foegen, A., and Stecker, P. M. (2009-2012). *Professional Development for Algebra Progress Monitoring. (Award # R324A090295)*. [Grant]. Washington, DC: National Center for Special Education Research.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Stecker, P. M. (1994). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *Am. Educ. Res. J.* 28, 617–641. doi: 10.2307/1163151
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Stecker, P. M. (2021). Bringing data-based individualization to scale: A call for the next generation technology of teacher supports. *J. Learn. Disabil.* 54, 319–333. doi: 10.1177/0022219420950654

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The opinions expressed here are those of the authors and do not represent views of the Institute or the United States Department of Education.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2022.944836/full#supplementary-material>

- Genareo, V. R., Foegen, A., Dougherty, B., DeLeeuw, W., Olson, J., and Karaman Dundar, R. (2019). Technical adequacy of procedural and conceptual assessment measures in high school algebra. *Assess. Effect. Interv.* 46, 121–131. doi: 10.1177/1534508419862025
- Helwig, R., Anderson, L., and Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *J. Special Educ.* 36, 102–112.
- Jung, P.-G., McMaster, K. L., Kunkel, A. K., Shin, J., and Stecker, P. M. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-analysis. *Learn. Disabil. Res. Pract.* 33, 144–155. doi: 10.1111/ldrp.12172
- Ketterlin-Geller, L. R., Gifford, D. B., and Perry, L. (2015). Measuring middle school students' algebra readiness: Examining validity evidence for three experimental measures. *Assess. Effect. Interv.* 41, 28–40.
- Kruzich, L. (2013). Thinkspace technology improves critical thinking and problem solving in simulations. *J. Acad. Nutr. Dietetics* 113:A68. doi: 10.1016/j.jand.2013.06.239
- Stecker, P. M., Fuchs, L. S., and Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychol. Sch.* 42, 795–819. doi: 10.1002/pits.20113
- Wagner, D. L., Hammerschmidt-Snidarich, S. M., Espin, C. A., Seifert, K., and McMaster, K. L. (2017). Pre-service teachers' interpretation of CBM progress monitoring data. *Learn. Disabil. Res. Pract.* 32, 22–31. doi: 10.1111/ldrp.12125
- Wolff, N., Johnson, J., Jones, S., Bender, H., Madeka, K., and Gahn, S. (2017). Eliminating writer's block: Flipped classroom meets ThinkSpace. *J. Acad. Nutr. Dietetics* 117:A69. doi: 10.1016/j.jand.2017.06.221



OPEN ACCESS

EDITED BY
Stefan Blumenthal,
University of Rostock, Germany

REVIEWED BY
Quan Zhang,
Jiaxing University, China
David Scheer,
Ludwigsburg University of Education,
Germany

*CORRESPONDENCE
Leanne R. Ketterlin-Geller
lkgeller@smu.edu

SPECIALTY SECTION
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

RECEIVED 10 May 2022
ACCEPTED 26 July 2022
PUBLISHED 26 September 2022

CITATION
Ketterlin-Geller LR, Sparks A and
McMurrer J (2022) Developing
progress monitoring measures: Parallel
test construction from the item-up.
Front. Educ. 7:940994.
doi: 10.3389/feduc.2022.940994

COPYRIGHT
© 2022 Ketterlin-Geller, Sparks and
McMurrer. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Developing progress monitoring measures: Parallel test construction from the item-up

Leanne R. Ketterlin-Geller^{1*}, Anthony Sparks² and
Jennifer McMurrer³

¹Education Policy and Leadership, Southern Methodist University, Dallas, TX, United States,
²American Medical Technologists, Chicago, IL, United States, ³Research in Mathematics Education,
Southern Methodist University, Dallas, TX, United States

Progress monitoring is a process of collecting ongoing samples of student work and tracking performance of individual students over time. Progress monitoring involves administering parallel sets of items to the same student on a regular basis (at least monthly) that are sensitive to changes in the student's understanding based on instruction. The sets of items administered over time should be parallel in difficulty so that differences in performance can be attributed to differences in the student's understanding as opposed to variability in item difficulty across sets. In this manuscript, we describe an approach to designing items that controls item-level variability by constraining the item features that may elicit different cognitive processing. This approach adapts the principles of Automated Item Generation (AIG) and includes carefully designing test specifications, isolating specific components of the content that will be assessed, creating item models to serve as templates, duplicating the templates to create parallel item clones, and verifying that the duplicated item clones align with the original item model. An example from an operational progress monitoring system for mathematics in Kindergarten through Grade 6 is used to illustrate the process. We also propose future studies to empirically evaluate the assertion of parallel form difficulty.

KEYWORDS

progress monitoring (PM) measures, mathematics education, computational fluency, instructional decision making, curriculum based measures

Introduction

Multi-tiered systems of support (MTSS) and data-based individualization (DBI) represent systems-level frameworks in which instruction and assessment are integrated into one coherent system with the goal of supporting positive outcomes for all students. These frameworks provide systematic approaches to link assessment results with classroom-level decisions to better align instruction with students' needs (Choi et al., 2017). Data from different assessments (e.g., universal screeners, diagnostic assessments, progress monitoring measures) are associated with specific instructional decisions so

as to provide teachers with guidance for interpreting student performance. As data are interpreted and teachers make decisions, they implement tiered instruction (e.g., Tier 1 Core Instruction, Tier 2 Intervention, Tier 3 Intensive Intervention) using evidence-based practices. As a result of implementing MTSS and DBI, teachers align students' learning needs as evidenced by assessment results with evidence-based instructional practices to support positive outcomes for all students (Powell et al., 2021).

A key decision underlying MTSS and DBI is determining whether students are making adequate progress to reach their learning goals (Ketterlin-Geller et al., 2019). The importance of this decision cannot be overstated because it serves as the key lever for changing students' instructional opportunities. If students are not making adequate progress toward their learning goals, it is incumbent on teachers to responsively change their instruction to better align with students' learning needs. Continually monitoring students' progress during the learning process provides teachers with the data they need to make these decisions.

The progress monitoring process

In a typical mathematics classroom, teachers use various approaches to monitor student learning including gathering data from both formal (e.g., quizzes, projects) and informal (e.g., questioning, noticing) sources. These data serve many purposes within the instructional decision-making framework such as identifying students' prior knowledge, understanding students' reasoning, or examining their flexibility using various representations or knowledge forms. Although these data help teachers understand student learning, they have limited utility for formally monitoring progress.

Within MTSS and DBI, formally monitoring progress refers to a systematic process of collecting ongoing samples of student work and tracking performance of individual students over time. The student's prior performance serves as the reference point for evaluating changes in understanding. The student's work samples must be taken from item sets that are administered over time. These item sets—sometimes referred to as *progress monitoring probes*—may take on different forms (such as reading passages, sentence completion), but in mathematics, they typically resemble a traditional test with items arranged in rows and columns on one or more pieces of paper. To monitor progress over time, teachers need approximately 20 probes that all measure the same construct and are of comparable difficulty so that changes in performance can be attributed to changes in student understanding, as opposed to variability in item difficulty. These concepts grew out of the work on curriculum-based measurement (CBM; Deno, 2003).

Research and development work on CBM as an approach to monitoring progress in mathematics began over 35 years

ago, and has evolved considerably over the years (c.f., Fuchs, 2004; Dawes et al., 2022). Although a large concentration of work has been done in elementary grades, CBMs have extended into early grades mathematics (c.f., Fuchs et al., 2007; Clarke et al., 2008) and secondary mathematics (c.f., Foegen et al., 2008). Mathematics CBMs most often measure grade-level computational fluency expectations, but some progress monitoring systems also include measures of students' conceptual understanding and application (Foegen et al., 2007). Recent research has explored the use of single-skill computational fluency measures (c.f., VanDerHeyden and Broussard, 2021; Dawes et al., 2022), yet more research is needed to determine whether this approach provides meaningful progress monitoring data over time (Fuchs, 2004). It follows that the assessed content of many mathematics CBMs may not represent the full depth and breadth of the grade-level content standards; however, the assessed content should be predictive of future outcomes and sensitive to small changes in students' understanding. To facilitate progress monitoring decisions, mathematics CBMs should be quick and easy to administer, efficient to score, and be psychometrically sound (Fuchs, 2004). Progress monitoring systems available from vendors, universities, or other resources have different characteristics and features so the probes are only considered parallel if they originate from within one progress monitoring system.

Tracking performance over time involves frequent administration of progress monitoring probes and graphing individual student's data. The most common administration frequency is weekly or every-other week, and no less frequently than monthly (Gersten et al., 2009). To accommodate this frequency within a school year, progress monitoring systems need to have at least 20 parallel forms. A comprehensive description of the data analysis and interpretation process is outside the scope of this manuscript. In brief, data are typically organized graphically for each individual student after multiple progress monitoring probes have been administered and teachers have a sufficient number of data points for making reliable interpretations. The slope of the line is interpreted as the student's observed rate of growth. This rate is compared to a goal rate that is typically established using published growth rates and the student's baseline score (see Jung et al., 2018 for research on the outcomes of different decision-making rules). Because the student's own performance serves as the interpretive lens for evaluating change over time, progress monitoring emphasizes growth and may facilitate positive associations between effort and outcome.

Creating parallel progress monitoring measures

As we have emphasized, multiple parallel forms of the same construct are needed to monitor individual student's

progress over time. Historically, parallel forms have been created and evaluated in one of two ways: (1) placing similar items on forms that are statistically compared for consistency (e.g., parallel form reliability), or (2) creating calibrated item banks such that all possible items are on the same scale (e.g., computerized adaptive tests [CAT]). Both of these approaches result in comparable scores across progress monitoring probes; however, methodological issues and inconsistent content may compromise the validity and reliability of these approaches.

Methodologically, these approaches require sufficiently large validation studies to evaluate the comparability of the forms and items. When statistically comparing forms for consistency, there are two common methods. First, the reliability of the alternate forms can be evaluated using a Classical Test Theory approach. Each of the 20 forms are administered to the same sample of students who are representative of the target population. Cross-correlation matrices are generated to evaluate the reliability of each parallel form. Second, a statistical method can be used to create statistically parallel forms of the same test called equating, which transforms raw scores to scale scores that are comparable (Kolen and Brennan, 2014). Equating is a process that results in interchangeable scores across multiple forms by statistically adjusting the scale so that the scores from each form have the same meaning when interpreted (American Educational Research Association [AERA] et al., 2014). One approach to equating is called common-subject equating, and uses a similar method as was described for calculating parallel-form reliability in which each of the 20 forms are administered to the same sample of students (Kolen and Brennan, 2014). Data from these students are used to adjust for differences in difficulty found across forms. Although this is a viable approach for equating parallel forms of some tests, given the 20 forms needed for progress monitoring systems, these designs place a burden on the students participating in the study.

Another way to create progress monitoring systems with multiple parallel forms is by using a calibrated item bank, such as a CAT. To create a CAT, items are typically calibrated using item response theory (IRT) modeling. Hundreds of items are needed to create an item bank sufficiently wide to reliably measure students with a range of ability levels and to administer 20 parallel forms without repeated exposure of the same item. All items need to be field tested using an equating designs so as to place all items on the same scale. Depending on the number of parameters being estimated, each item requires 250–1,000 responses for accurate calibration (Rupp, 2003). Given the large sample of students needed to calibrate the large set of items, the costs and timeliness of this approach may be prohibitive. As such, methodological issues limit the feasibility of these approaches for creating parallel forms within a progress monitoring system.

In addition to methodological issues associated with statistical approaches to evaluating comparability of parallel

forms, the underlying assumptions of content comparability may not be tenable. To support valid decision-making regarding students' progress, data should facilitate inferences about students' growth on consistently measured content standards. If the content of the progress monitoring measures changes over time, students may perform differently across forms for reasons that are not necessarily related to learning the targeted knowledge and skills. Two salient issues emerge: (1) items may have similar difficulty statistics (e.g., p -values, item difficulty parameters) but cover different content, and (2) content differences may differentially impact students' responding behaviors based on prior knowledge, exposure or opportunity to learn the content, fluency across number ranges and systems, etc. These differences may lead to increases or decreases in students' scores on the progress monitoring probes that do not accurately reflect changes in understanding. As such, teachers' interpretations of growth (or lack thereof) may be inaccurate, thereby jeopardizing the validity of their decision making.

To illustrate the challenges of using item difficulty statistics to evaluate comparability, consider the released items from the Grade 4 National Assessment of Educational Progress (NAEP) in Mathematics administered in 2017 presented in Figure 1. Figure 1 displays two items from the "Number properties and operations" domain within NAEP and the estimate of item difficulty expressed as p -values (proportion of students responding correctly to the total number of respondents). Item 1 requires students to solve a multi-step problem in context. Item 2 focuses on place value understanding, and assesses students' ability to identify the number represented by a set of based-ten blocks. Even though these items assess the same mathematical domain and the p -values indicate comparable difficulty, they measure different mathematical content that may elicit different levels of cognitive engagement that impact individual student's responding behaviors. As such, aggregated statistics might mask differences in individual student's performance. In instances where these statistics are used to determine form comparability for progress monitoring probes, students may perform differently across forms that is not due to growth.

Even when content is held constant, subtle differences in wording or students' opportunity to learn the content may impact item difficulty. To illustrate these issues, consider the following released items from the Grade 4 NAEP in Mathematics that are designed to assess students' ability to use place value to determine the amount of increase or decrease in whole numbers. Figure 2 includes two items and their respective p -values as reported by NAEP. Both items require students to identify by how much a given number would increase if the value of a specific digit were changed. Item 1 was considerably less difficulty than Item 2 in that 62% of the respondents answered correctly as compared to 36% for Item 2. For Item 1, the distractors were selected roughly

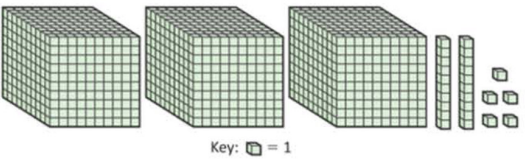

No	Item	p-value
1	<p>Mr. Franklin wants to buy an eraser for every fourth-grade student.</p> <p>There are 12 erasers in each box.</p> <p>There are 7 fourth-grade classes with 24 students in each class.</p> <p>How many boxes of erasers does Mr. Franklin need to buy?</p> <p> <input type="radio"/> A 2 <input type="radio"/> B 14 <input type="radio"/> C 43 <input type="radio"/> D 84 </p> <p>Clear Answer</p>	0.43
2	 <p>Key:  = 1</p> <p>Which of the following numbers is represented by the base ten blocks?</p> <p> <input type="radio"/> A 325 <input type="radio"/> B 370 <input type="radio"/> C 3,025 <input type="radio"/> D 3,205 </p> <p>Clear Answer</p>	0.40

FIGURE 1

NAEP items with similar difficulty assessing number properties and operations. Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2017 Mathematics Assessment.

No	Item	p-value
1	<p>By how much will the value of the number 4,372 increase if the 3 is replaced with a 9 ?</p> <p>A. 6</p> <p>B. 60</p> <p>C. 600</p> <p>D. 6,000</p>	0.62
2	<p>By how much would 217 be increased if the digit 1 were replaced by a digit 5?</p> <p>A. 4</p> <p>B. 40</p> <p>C. 44</p> <p>D. 400</p>	0.36

FIGURE 2

NAEP items with different difficulty assessing place value. Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2007 and 1992 Mathematics Assessments.

equally with less than 15% of the respondents selecting any one distractor; Distractor B was selected the least at 8%. For Item 2, Distractors A and C were selected by 22% of the respondents, and Distractor D was selected by 14% (6% of the respondents omitted this question).

On the surface, these items appear to be very similar in content and structure. Subtle differences in the wording of the stem may have caused differences in the item difficulty estimates as might have students' opportunity to learn these place-value concepts. Item 1 was operational in 2007, whereas Item 2 was operational in 1992. Given the change in content standards and expectations in the 25-year span between administrations, it is likely that changing curricular expectations impacted students' exposure to these concepts. As this example points out, factors other than the content and structure of an item may impact item difficulty.

An emerging approach to creating parallel forms is using automated item generation (AIG). The goal of AIG is "to produce large numbers of high-quality items that require little human review prior to administration" (Alves et al., 2010, p. 2). Two approaches emerge within the AIG framework: automatic and semi-automatic (Prasetyo et al., 2020). Automatic AIG incorporates the use of natural language processing for the generation of questions, answers, and distractors simultaneously. For semi-automatic AIG, an expert develops a stem of questions that can be adapted to create new items. These new items can either be clones or variants. Clones are similar items with comparable psychometric properties, while variants possess different psychometric properties. Semi-automatic AIG is primarily contained in three steps (Royal et al., 2018). First, content experts create a cognitive map that identifies the content for inclusion in the assessment; this serves as the assessment blueprint. Next, the experts develop a template or item model for the content. Lastly, a computer algorithm combines various elements of content provided by the experts to generate new items. In the context of creating progress monitoring measures, AIG holds promise for generating parallel forms; however, limited application of this technology exists in this context.

In this manuscript, we describe an application of the principles of semi-automatic AIG to create a progress monitoring system in mathematics for students in Kindergarten through Grade 6. Each grade included 20 parallel forms of 30–40 items on each form. The project described here followed the general framework of semi-automatic AIG but employed only humans in the development process. After specifying the test blueprint, we created item templates that constrained the test and item specification to isolate specific components of the content that would be assessed by each item. These templates were then used by item writers to create item clones for 20 parallel forms. The goal for using this approach is to support the inference that students engage with items on parallel

forms in comparable ways that are related to their present-level of understanding.

At present, the items created through this process have not been psychometrically evaluated to substantiate the claim that they are parallel in difficulty. Once field testing data are available, these sources of evidence can be combined to examine the claim that the progress monitoring system can be used to monitor growth in individual student's understanding.

Illustrative example of mathematics progress monitoring

The purpose of the progress monitoring system described in this manuscript was to facilitate educators' decisions about students' growth in the computations-based content standards in Kindergarten (K) through Grade 6. Results from multiple administrations of these probes would allow educators to make inferences about changes in individual student's computational fluency on grade-level content standards. For the remainder of this manuscript, we will refer to this project as the COMP-PM.

The following description illustrates the three phases of the semi-automated AIG framework, as applied to develop the COMP-PM: (1) specify the test specifications and blueprint, (2) develop the template for item clones, and (3) automate the item cloning process. We also present the validity evidence we collected to evaluate the assumptions that the items are clones and will result in parallel forms.

Phase 1: Specify the test specifications and blueprint

Our first step in developing parallel progress monitoring forms for the COMP-PM was to determine the test specifications. Test specifications are intended to articulate multiple aspects of the operational test including the item and test format, number of items, scoring rules, interpretive reference, and time limits, and should be based on the intended interpretations and uses of the test results (American Educational Research Association [AERA] et al., 2014). Subsumed within the test specification is the test blueprint, which details the content covered at the test- and item-level.

To begin, we determined the computations-based content standards that would be assessed on the progress monitoring measures. For most grades, these standards were clearly specified as fluency-based expectation. For example, computational fluency is clearly expressed in the Grade 2 standard: Students are expected to recall basic facts to add and subtract within 20 with automaticity. However, for some grades, the computations-based expectations were intertwined with other content standards. In Grade 3, for

example, students are expected to use strategies and algorithms, including the standard algorithm, to multiply a two-digit number by a one-digit number. Computational fluency expectations are expressed as students are expected to use algorithms to multiply.

To identify the assessable content for the COMP-PM, two experts in mathematics education closely examined the state content standards to identify the computations-based expectations by grade. The experts reviewed the content standards to pinpoint individual skills that related to computational fluency, and importantly, identified the number range in which those skills would be applied. For Kindergarten and Grade 1, two early numeracy constructs were selected because of the predictive evidence with future mathematics performance. A third expert reviewed the final list of assessable content; any disagreements were discussed until consensus was reached.

Next, key decisions related to the format of the operational test were made using prior research on the design of CBM (c.f., Fuchs and Fuchs, 1997; Foegen et al., 2007). These decisions included:

- **Item format:** Items are formatted as constructed response to allow students to directly demonstrate their knowledge and skills. Depending on the grade and alignment with the content standards, items will be presented horizontally, vertically, or both.
- **Test format:** Forms are created to allow ample room for students to solve and record their response to each item. In Kindergarten and Grade 1, forms are presented horizontally to maximize space; each subtest is formatted as a separate form. For grades 2–6, all items are formatted vertically as one operational form with 30 items arranged in six rows of five items each. Item arrangement is intentional to vary the placement of items by content representation and difficulty. Item difficulty will mirror a normal distribution.
- **Number of subtests and items per subtest:** For Kindergarten and Grade 1, two subtests each with 20 items are needed to assess the selected content standards.
- **Scoring rules:** All items are scored dichotomously to minimize scoring time and errors.
- **Interpretive reference:** Consistent with other progress monitoring systems, scores on the COMP-PM will be interpreted in relation to the student's prior performance. As such, no criteria or normative data are provided to aid in interpretation.
- **Time limits:** The time constraints for administration are needed to maximize students' opportunities to demonstrate their knowledge while still minimizing the impact of administration on instructional time. Administration is standardized across parallel forms so that students always have the same amount of time.

Grade	Administration Time
K	1 min each side
1	1 min each side
2	A total of 2 min
3	A total of 2 min
4	A total of 4 min
5	A total of 4 min
6	A total of 4 min

Using the assessable content and the test specifications, we created a generalize test blueprint to identify the number of items needed to assess each skill. The number of items associated with each content standard was determined based on the relative importance and priority of the skill within the grade. [Figure 3](#) illustrates the test blueprint for Grade 5.

At the end of Phase 1, we had fully articulated the test specifications for the operational progress monitoring system, and detailed the content to be assessed. The test specifications and test blueprint were reviewed by mathematics education experts at the state education agency. Iterative refinements were made based on their feedback.

Phase 2: Develop the template for item clones

The next phase focused on creating the item templates from which item clones would be generated. A unique item template was needed for each of the 30–40 items per grade. Item templates isolate specific components of the content that are assessed by each item. The purpose of the item template is to specify (and thereby constrain) as many factors as possible that could cause students to engage with the items using different cognitive processes. To the extent that these cognitive processes change the elicited knowledge and skills, the resulting items may not be clones. The goal of this phase was to create 20 clones for each of the 30–40 items per grade so that the resulting 20 forms would be parallel in both structure and content, with the intention of being comparable in difficulty.

To begin, we created a fine-grained content matrix that specified the detailed content that would be assessed by each item. During this step, we dissected multi-component content standards into subcomponents that could be a source of variability in the items. For example, a Grade 5 content standard specifies that students can multiply with fluency a three-digit number by a two-digit number using the standard algorithm. Variability in the value of the three-digit and two-digit numbers may impact the difficulty of these items. As such, for the fine-grained content matrix, we specified which multiplicand was a multiple of ten. [Figure 4](#) illustrates the content matrix for Grade 5.

Standard and subcomponents		Total Number of Items
Multiply with fluency a three-digit number by a two-digit number using the standard algorithm.		5
Solve with proficiency for quotients of up to a four-digit dividend by a two-digit divisor using strategies and the standard algorithm.		5
Solve for products of decimals to the hundredths, including situations involving money, using strategies based on place-value understandings, properties of operations, and the relationship to the multiplication of whole numbers.		3
Solve for quotients of decimals to the hundredths, up to four-digit dividends and two-digit whole number divisors, using strategies and algorithms, including the standard algorithm.		3
Add and subtract positive rational numbers fluently.		10
Add positive rational numbers	5	
Subtract positive rational numbers	5	
Divide whole numbers by unit fractions and unit fractions by whole numbers.		4
Divide whole numbers by unit fractions	2	
Divide unit fractions by whole numbers	2	
Total		30

FIGURE 3
Blueprint for Grade 5.

Also during this step, we assigned a specific number of items to have low, medium, and high difficulty so as to include a range of difficulty levels within the form. Difficulty was determined by several characteristics, including the specific numbers included, the number of steps needed to complete the problem, the amount of information that needed to be retained in working memory, and the number of components needed to execute the algorithm.

To facilitate creating the unique item templates, we created a form that included an algebraic representation of the item in addition to constraints on the item to keep the difficulty consistent across item clones. We also included space to explore common misconceptions in the solving of the problem. The purpose of including misconceptions was to capture common misunderstandings students have for each concept. Data from misconceptions may also provide diagnostic information in the future. Misconceptions were drawn from literature and item writers' teaching experience. [Figure 5](#) shows the item template.

We convened two meetings with 24 content-area experts (e.g., teachers, instructional coaches) to develop 220 item models across Kindergarten through Grade 6. Content area experts were recruited from professional networks with local school districts. Qualifications included:

- Bachelor's degree or higher in mathematics, education, or related field
- Three years teaching experience in the state in Grade(s) K-6
- Deep understanding of the state content standards
- Ability to accept and incorporate critical feedback

- Proficiency in Microsoft Word/Excel
- Ability to scan/upload files to an online repository
- Ability to adhere to tight timelines
- Experience with writing mathematics assessment items in Grades K-6 (preferred)
- Extensive background in supporting elementary or middle school teachers as a mathematics coach (preferred)

The purpose of these meetings was to train the content-area experts on the purpose and procedures for creating each item template and the corresponding item clones, and create all item templates from which the item clones would be created at a later date. During the meetings, we provided background information on progress monitoring, plausible misconceptions and errors, and factors that impact item difficulty. We also reviewed the test blueprint and content matrix for each grade. Then, we discussed item writing procedures and reviewed the completed item template and three sample item clones presented in [Figure 6](#). We used the item template in [Figure 6](#) to illustrate the importance of specifying the misconceptions and being exhaustive in the constraints to support writing item clones. For Item Clone 2, responses to Misconception 2 and 3 lead to the same answer. To provide diagnostically relevant information, the misconceptions should lead to different answers. For Item Clone 3, the response is only two digits, which may impact students' cognitive processing. This led to a discussion about the sufficiency of the original constraints, and resulted in updating the constraints to specify that $a > d + 1$.

Content Standards and Subcomponents	# of Items	Relative Item Difficulty		
		Low (n=12)	Medium (n=12)	High (n=6)
Multiply with fluency a three-digit number by a two-digit number using the standard algorithm. < 100,000	5	2	2	1
Multiply a three-digit number by a two-digit number (both multiplicands are multiples of ten within 500 & 99)	1			
Multiply a three-digit multiple of ten by a two-digit number (within 500 & 99)	1			
Multiply a three-digit number by a two-digit multiple of ten (within 500 & 99)	1			
Multiply a three-digit number by a two-digit number, neither a multiple of ten ([1] within 500 & 99 [2] within 999 & 99)	2			
Solve with proficiency for quotients of up to a four-digit dividend by a two-digit divisor using strategies and the standard algorithm. < 100,000	5	2	2	1
for quotients of a four-digit dividend by a two-digit divisor ([1] within 5000 & 99 [2] within 9999 & 99)	2			
for quotients of a three-digit dividend by a two-digit number ([1] within 500 & 99 [2] within 999 & 99)	2			
for quotients of a two-digit dividend by a two-digit divisor (within 99 & 50)	1			
Solve for products of decimals to the hundredths, including situation involving money, using strategies based on place-value understandings, properties of operations, and the relationship to the multiplication of whole numbers. < 100,000	3	1	1	1
for products of decimals to the tenth (within 99.99 & 50)	1			
for products of decimals to the hundredths ([1] within 555.99 & 50 & 99 [2] within 999.99 & 99)	2			
Solve for quotients for decimals to the hundredths, up to four-digit dividends and two-digit whole number divisors, using strategies and algorithms, including the standard algorithm. < 100,000	3	2	1	0
for quotients of decimals to the hundredths, four-digit dividends and two-digit whole number divisors (within 5555.99 & 99)	1			
for quotients of decimals to the hundredths, three-digit dividends and two-digit whole number divisors (within 555.99 & 99)	1			
for quotients of decimals to the hundredths, two-digit dividends and two-digit whole number divisors (within 99.99 & 99)	1			
Add and subtract positive rational numbers fluently. < 100,000	10	4	4	2
Add positive rational numbers	5	2	2	1
positive rational numbers with like denominators (denominators within [1] 999 [2] 9999)	2			
positive rational numbers with different denominators (denominators within [1] 999 [2] 9999)	2			
positive rational numbers represented as decimals (decimals within thousandths)	1			
Subtract positive rational numbers	5	2	2	1
positive rational numbers with like denominators (denominators within [1] 999 [2] 9999)	2			
positive rational numbers with different denominators (denominators within [1] 999 [2] 9999)	2			
positive rational numbers represented as decimals (decimals within thousandths)	1			
Divide whole numbers by unit fractions and unit fractions by whole numbers. < 100,000	4	1	2	1
Divide whole numbers by unit fractions	2			
whole number multiple of ten by unit fraction (whole number within 9999 & denominator within 999)	1			
whole number non-multiple of ten by a unit fraction (whole number within 999 & denominator within 99)	1			
Divide unit fractions by whole numbers	2			
unit fraction by whole number multiple of ten (Denominator within 9999 & whole number within 999)	1			
unit fraction by whole number non-multiple of ten (Denominator within 999 & whole number within 99)	1			
Total	30	12	12	6

FIGURE 4
Content matrix for Grade 5.

After the initial group discussion, content-area experts were divided into grade-level groups to write item templates. As item templates were completed, they were evaluated by two project team members and other content-area experts through an extensive and systematic process. The primary review criteria included alignment with the test blueprint and content matrix,

sufficiency of the constraints to maintain item difficulty, and plausibility of the misconception. The Item Template was also reviewed for alignment with the proposed difficulty level so as to ensure distribution of item difficulties as specified in the content matrix. Where needed, we modified items to maintain item difficulties across the distribution. In instances where the

Algebraic Form of Item:	Example Item 1:	Responses	
		Correct Response:	
Item Constraints:	Misconception 1:	Alternate Response 1:	
	Misconception 2:	Alternate Response 2:	
	Misconception 3:	Alternate Response 3:	
Item/Constraint/Error Review			Final Responses
Item 2: Correct Response: Alternate Response 1: Alternate Response 2:	Reviewer's Feedback: (Initials)	RME Approval: (Initials)	Correct:
			Alt 1:
			Alt 2:
			Alt 3:
Item 3: Correct Response: Alternate Response 1: Alternate Response 2:	Reviewer's Feedback: (Initials)	RME Approval: (Initials)	Correct:
			Alt 1:
			Alt 2:
			Alt 3:

FIGURE 5
Item model template.

two reviewers disagreed with the difficulty rating, the reviewers discussed until consensus was reached.

As part of this initial review, the first three item clones (labeled Item 1, 2, 3 in the Item Template in [Figure 5](#)) were carefully examined. Each item clone was evaluated to verify that it matched the constraints specified the Item Template and elicited the same cognitive processes as the other item clones. If the cognitive processes varied across clones, the clone was modified to align with the constraints. If the constraints were met but the item clone still elicited different cognitive processes, the constraints in the Item Template were updated to better control for variability in the cognitive processes. Some item templates required multiple rounds of revision before being finalized.

To gather content-related validity evidence, the finalized item templates were reviewed by five external reviewers with expertise in mathematics and special education, with a particular emphasis on progress monitoring and/or curriculum based measurement. Four of the external reviewers reviewed the item templates for one grade; one external reviewer reviewed the item templates for two grades. Qualifications to serve as an external reviewer included:

- A doctoral degree in mathematics, education, or related field;

- Five years of experience working in a teaching, administrative, or university setting in their field;
- A deep understanding of mathematics content standards;
- Experience with writing mathematics assessment items in Grades K-6; and
- Extensive background in supporting elementary or middle school teachers, preferred.

During the external review process, each item template was reviewed for alignment with the test blueprint and content matrix, alignment with the difficulty level, feasibility and sufficiency of the constraints, comparability of cognitive processing, plausibility of the misconception, and likelihood of generating 20 alternate forms. External reviewers provided feedback for each criteria using a four-point Likert scale (**1: strongly disagree, 2: disagree, 3: agree, 4: strongly agree**). For any criteria that received a rating of 1 or 2, we requested written rationale for their rating and recommendations to help improve the item template.

Table 1 describes the percent agreement of the external reviewers' ratings across grades for each criteria. Experts agreed or strongly agreed that 77–100% of the item templates aligned with the content standards. Alignment to the assigned difficulty agreement ranged from 57 to 95%. Agreement that the item constraints would yield 20 comparable items ranged from 77 to

Filled Item Model Template

TIER Item Writing Template			
TEKS Standard:			
Algebraic Form of Item:	$c > f$	Example Item 1:	Responses
$ab.c$ $- de.f$	$b < e, b > 0$ $a > d, b \neq 0$	65.8 $- 39.7$	Correct Response: 26.1
Item Constraints:		Misconception 1:	Alternate Response 1:
Regrouping in ones place		Disregard place value	261
No regrouping in tenths place		Misconception 2:	Alternate Response 2:
Minuend > 50		Regrouping error	36.1
Subtrahend < 50		Misconception 3:	Alternate Response 3:
		Subtraction not commutative	34.1
Item/Constraint/Error Review			Final Responses
Item 2:	73.6 $- 28.4$	Reviewer's Feedback:	Correct:
Correct Response:		RME Approval:	Alt 1:
Alternate Response 1:		(Initials)	Alt 2:
Alternate Response 2:		(Initials)	Alt 3:
Item 3:	51.7 $- 47.3$	Reviewer's Feedback:	Correct:
$a > d + 1$		RME Approval:	Alt 1:
		(Initials)	Alt 2:
		(Initials)	Alt 3:

FIGURE 6
Filled item model template.

100%. Agreement of the appropriateness of the misconceptions ranged from 44 to 100% and agreement in the appropriateness of the alternate responses ranged from 50 to 100%. The criteria with the lowest level of agreement was misconceptions.

Using the external reviewers' rationale and recommendations for improvements, at least one project team member reviewed and revised the item templates that received a rating of 1 or 2 (strongly disagree or disagree). An independent reviewer from the project team served as a verifier; this team member reviewed the external reviewer's feedback and the revision to verify that the issue was adequately addressed.

TABLE 1 External review percent agree/strong agree.

Criteria	K*	1*	2	3	4	5	6
Alignment to content standards	100%	100%	77%	90%	93%	97%	100%
Difficulty alignment	93%	95%	77%	90%	73%	57%	77%
Constraints	100%	98%	80%	100%	87%	77%	83%
Comparable forms	0%	88%	100%	100%	83%	100%	100%
Misconceptions	44%	100%	100%	100%	100%	100%	100%
Alternate responses	58%	100%	97%	97%	50%	50%	90%

*Not all items had misconceptions/alternate responses.

Any discrepancies were reconciled with the original project team member and/or the external reviewer.

At the conclusion of Phase 2, we had 220 unique item templates across Kindergarten through Grade 6. Through extensive and systematic internal and external review processes, we reviewed and revised the item templates to verify that they met the criteria. Content-related validity evidence supported our claim that the item templates measured the content specified in the content matrix and item clones would be comparable in difficulty and elicit similar cognitive processes. As a result, the item templates were used to initiate Phase 3 in which the item clones would be created.

Phase 3: Automate the item cloning process

The purpose of Phase 3 was to create 20 item clones for each of the 220 item templates. As previously noted, each item template included the algebraic form of the item, constraints to maintain comparability of content, and cognitive processes, possible misconceptions, and the corresponding alternate responses, and three sample item clones. Constraining

input to finalize the item clones and submit for final review by the project team. In some cases, multiple iterations of revisions were needed before the completed set of 20 item clones was approved.

At the conclusion of Phase 3, 4,180 item clones were finalized for 220 item templates to be distributed across 30–40 operational forms for Kindergarten through Grade 6. To aid in the placement of the items in the operational forms, we created a form blueprint that aligned with the test specifications presented earlier. The form blueprint is a schematic that documents where the items are to be placed on the operational forms. Items assessing similar content standards were dispersed across the form. Item difficulty was also considered when distributing the items and mirrored a normal distribution; the number of most difficult items was greatest in the middle of the form. The first row of items on every form did not include any of the most difficult items. This placement was intentional to allow students with varying ability levels to demonstrate their knowledge, skills, and abilities, and was intended to minimize anxiety.

Once the items were placed, the final forms across all grades were reviewed. Item formatting was examined and content cueing was considered to make sure students' responses to one item would not influence their responses to others. Once these forms were finalized, they were used to create the final answer key and student forms.

Discussion

The current paper describes the process of developing a progress monitoring system in mathematics for students in Kindergarten through Grade 6. We adapted a semi-automatic

Cousin Item 1		Final Responses	
<p>Peer Review Feedback Checklist</p> <p>Alignment</p> <ul style="list-style-type: none"> ○ Algebraic Form ○ Constraints ○ Difficulty <p>Accuracy of Math</p> <ul style="list-style-type: none"> ○ Item ○ Alternate responses ○ 20 items total (unless otherwise noted) <p>NOTES:</p> <p>(Initials)</p>	<p>RME Feedback</p> <p>(Initials)</p>	Correct:	
		Alt 1:	
		Alt 2:	
		Alt 3:	
		ITEM FINALIZED	
		DATE	Initials

Development process for writing and verifying item clones

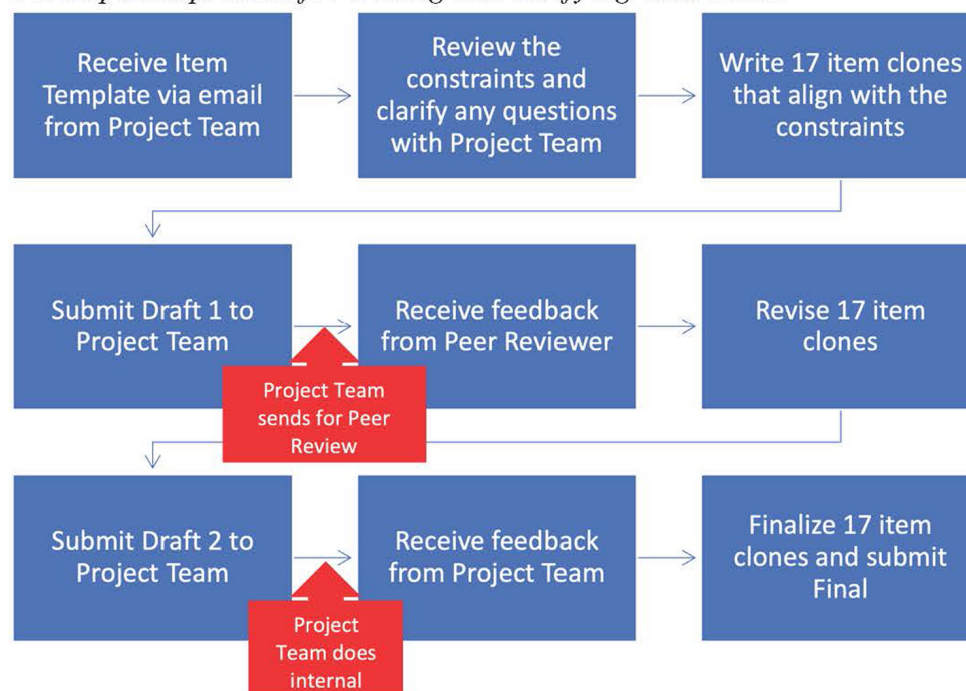


FIGURE 8
Development process for writing and verifying item clones.

item generation (AIG) approach to address methodological and content-related issues with traditional approaches to developing and validating progress monitoring systems. Using AIG, items are developed based on an item template that controls for variability in item difficulty. Controlling for item-level variability is important in the development of progress monitoring tools, which depend on items of comparable difficulty across multiple forms. These comparable items allow stakeholders the ability to monitor individual student's progress across the administration of the different probes of the same construct.

In this manuscript, we describe the three phases of the adapted AIG approach that we implemented. Throughout each phase, we collected content-related evidence for validity and made iterative improvements. During Phase 1, to verify the alignment with the state content standards in mathematics, the test specifications and test blueprint were reviewed by mathematics education experts at the state education agency. In Phase 2, the finalized item templates were reviewed by five external reviewers with expertise in mathematics and special education, with a particular emphasis on progress monitoring and/or curriculum based measurement. They reviewed each item template for alignment with the test blueprint and content matrix, alignment with the intended difficulty level, feasibility, and sufficiency of the constraints, comparability of cognitive

processing, plausibility of the misconception, and likelihood of generating 20 alternate forms. Finally, in Phase 3, all of the item clones went through a rigorous internal review by content-area experts and mathematics education researchers. At each phase, the quality was assessed and revisions were made to improve the final items.

The approach described in the current paper does not take the place of pilot or field testing and empirical evaluation of the comparability of the forms. For the current research, we used multiple reviews from experts to support the assumption that item difficulty remained consistent across forms. However, a limitation of the current research is the absence of psychometric data to verify this assertion. For example, using pilot or field test data, we need to analyze the comparability of items across multiple forms to assess whether item difficulty is maintained. Analyses could include comparing item difficulty and discrimination parameters derived from IRT modeling or analyses based in classical test theory. The results of these analyses could help support the claim that these items measure the same construct of computational fluency at the same difficulty across forms.

Differences in difficulty across forms may be detected. In these instances, forms can be equated to adjust for differences in difficulty. To avoid the issues previously described with the common-subjects method, a viable method for equating

progress monitoring probes would be to embed a set of common items (also known as anchor items) across each of the 20 forms during pilot or field testing. Although a detailed description of equating designs is beyond the scope of this manuscript, equating *via* anchor items allows the forms to be administered to different samples of students (see [Hanson and Beguin, 2002](#) for a more detailed description of the common-item equating design). Prior to operationalizing the progress monitoring system, the anchor items should be removed from the forms. One implication for this approach to creating parallel progress monitoring probes is the resulting use of scale scores. To facilitate teachers' use and interpretation of progress monitoring data, raw scores are typically computed and graphed. Using scale scores would require teachers to use score conversion tables for each form, which may impact their implementation.

Conclusion

In conclusion, this manuscript demonstrated the value of using an adapted AIG process to facilitate rapid development a progress monitoring system in mathematics. Content-related validity evidence supported the claims that both content and structure of the items were consistent across forms. Additional empirical evidence is needed to substantiate these claims.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

References

- Alves, C. B., Gierl, M. J., and Lai, H. (2010). "Using automated item generation to promote principled test design and development," in *Paper Presented at the Annual Meeting of the American Educational Research Association* (Denver, CO).
- American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- Choi, J. H., Meisenheimer, J. M., McCart, A. B., and Sailor, W. (2017). Improving learning for all students through equity-based inclusive reform practices: Effectiveness of a fully integrated schoolwide model on reading and math achievement. *Remedial Spec. Educ.* 38, 28–41. doi: 10.1177/0741932516644054
- Clarke, B., Baker, S., Smolkowski, K., and Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement. *Remedial Spec. Educ.* 29, 46–57. doi: 10.1542/peds.2016-2651
- Dawes, J., Solomon, B., and McCleary, D. F. (2022). Precision of single-skill mathematics CBM: Group versus individual administration. *Assess. Effect. Interv.* 47, 170–178. doi: 10.1177/15345084211035055
- Deno, S. L. (2003). Developments in curriculum-based measurement. *J. Spec. Educ.* 37, 184–192. doi: 10.1177/00224669030370030801
- Foegen, A., Jiban, C., and Deno, S. (2007). Progress monitoring measures in mathematics. *J. Spec. Educ.* 41, 121–139. doi: 10.1177/00224669070410020101
- Foegen, A., Olson, J. R., and Impeccoven-Lind, L. (2008). Developing progress monitoring measures for secondary mathematics: An illustration in algebra. *Assess. Effect. Interv.* 33, 240–249. doi: 10.1177/1534508407313489
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychol. Rev.* 33, 188–192.
- Fuchs, L. S., and Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus Except. Child.* 30, 1–16. doi: 10.17161/fec.v30i3.6758
- Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., and Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Except. Child.* 73, 311–330. doi: 10.1177/001440290707300303
- Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., and Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Rev. Educ. Res.* 79, 1202–1242. doi: 10.3102/0034654309334431
- Hanson, B. A., and Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Appl. Psychol. Meas.* 26, 3–24. doi: 10.1177/0146621602026001001

Author contributions

LK-G conceived and designed the study and wrote the first draft of the manuscript. JM managed the implementation of the study. AS contributed to the implementation of the study and wrote sections of the manuscript. All authors contributed to manuscript revisions, read, and approved the submitted version.

Funding

This research reported was funded under a contract received by the University of Texas at Austin on behalf of the Texas Education Agency.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Jung, P.-G., McMaster, K. L., Kunkel, A. K., Shin, J., and Stecker, P. M. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-analysis. *Learn. Disabil. Res. Pract.* 33, 144–155. doi: 10.1111/ldrp.12172
- Ketterlin-Geller, L. R., Powell, S., Chard, D., and Perry, L. (2019). *Teaching Math in Middle School: Using MTSS to Meet All Students' Needs*. Baltimore, MD: Brookes Publishing.
- Kolen, M. J., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd Edn. New York, NY: Springer. doi: 10.1007/978-1-4939-0317-7
- Powell, S. R., Lembke, E., Ketterlin-Geller, L. R., Petscher, Y., Hwang, J., Bos, S. E., et al. (2021). Data-based individualization in mathematics to support middle-school teachers and their students with mathematics learning difficulty. *Stud. Educ. Eval.* 69:100897. doi: 10.1016/j.stueduc.2020.100897
- Prasetyo, S. E., Adjij, T. B., and Hidayah, I. (2020). "Automated item generation: model and development technique," in *Paper Presented at the 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)* (Semarang). doi: 10.1109/ICITACEE50144.2020.9239243
- Royal, K. D., Hedgpeth, M., Jeon, T., and Colford, C. M. (2018). Automated item generation: The future of medical education assessment? *Eur. Med. J.* 2, 83–93.
- Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for windows. *Int. J. Test.* 3, 365–384. doi: 10.1207/S15327574IJT0304_5
- VanDerHeyden, A. M., and Broussard, C. (2021). Construction and examination of math subskill mastery measures. *Assess. Effect. Interv.* 46, 188–196. doi: 10.1177/1534508419883947



OPEN ACCESS

EDITED BY

Erica Lembke,
University of Missouri, United States

REVIEWED BY

Kaiwen Man,
University of Alabama, United States
Francis O'Donnell,
National Board of Medical Examiners,
United States

*CORRESPONDENCE

Christine A. Espin
espinca@fsw.leidenuniv.nl

SPECIALTY SECTION

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

RECEIVED 15 April 2022

ACCEPTED 22 August 2022

PUBLISHED 30 September 2022

CITATION

van den Bosch RM, Espin CA,
Sikkema-de Jong MT, Chung S,
Boender PDM and Saab N (2022)
Teachers' visual inspection
of Curriculum-Based Measurement
progress graphs: An exploratory,
descriptive eye-tracking study.
Front. Educ. 7:921319.
doi: 10.3389/feduc.2022.921319

COPYRIGHT

© 2022 van den Bosch, Espin,
Sikkema-de Jong, Chung, Boender
and Saab. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Teachers' visual inspection of Curriculum-Based Measurement progress graphs: An exploratory, descriptive eye-tracking study

Roxette M. van den Bosch¹, Christine A. Espin^{1*},
Maria T. Sikkema-de Jong^{1,2}, Siuman Chung¹,
Priscilla D. M. Boender¹ and Nadira Saab³

¹Department of Education and Child Studies, Leiden University, Leiden, Netherlands, ²Leiden Institute for Brain and Cognition (LIBC), Leiden University, Leiden, Netherlands, ³Graduate School of Teaching (ICLON), Leiden University, Leiden, Netherlands

In this exploratory descriptive study, we use eye-tracking technology to examine teachers' visual inspection of Curriculum-Based Measurement (CBM) progress graphs. More specifically, we examined which elements of the graph received the most visual attention from teachers, and to what extent teachers viewed graph elements in a logical sequence. We also examined whether graph inspection patterns differed for teachers with higher- vs. lower-quality graph descriptions. Participants were 17 fifth- and sixth-grade teachers. Participants described two progress graphs while their eye-movements were registered. In addition, data were collected from an expert to provide a frame of reference for interpreting the teachers' eye-tracking data. Results revealed that, as a group, teachers devoted less visual attention to important graph elements and inspected the graph elements in a less logical sequence than did the expert, however, there was variability in teachers' patterns of graph inspection, and this variability was linked to teachers' abilities to describe the graphs. Directions for future studies and implications for practice are discussed.

KEYWORDS

progress monitoring, teachers, graph comprehension, eye-tracking, CBM

Introduction

Teachers are increasingly expected to use data to guide and improve their instructional decision-making. In general education, this data-use process often is referred to as *Data-Based or Data-Driven Decision Making* (e.g., see Mandinach, 2012; Schildkamp et al., 2012). In special education it is referred to as *Data-Based Instruction or Individualization* (e.g., see Kuchle et al., 2015; Jung et al., 2017). Despite differences in terminology, researchers in general and special education draw upon

similar data-use models, which typically include the following steps: (a) identify and define the problem; (b) collect and analyze data; (c) interpret/make sense of the data; (d) make an instructional decision (e.g., see Mandinach, 2012; Deno, 2013; Beck and Nunnaley, 2021; Vanlommel et al., 2021). It is not only the data-use models that are similar across general and special education, but also the concerns about teachers' ability to successfully implement the models, especially their ability to implement steps (c) and (d). In both general and special education, research has shown that teachers have difficulty interpreting data and making effective instructional decisions based on these interpretations (e.g., see Stecker et al., 2005; Datnow and Hubbard, 2016; Gleason et al., 2019; Espin et al., 2021a; Mandinach and Schildkamp, 2021).

Although it is clear from the research that teachers have difficulty interpreting data and making instructional decisions, it is not clear why teachers have such difficulties. Answering the why question requires an understanding of the processes underlying teachers' data-based decision making. In the current study, we examine the processes underlying teachers' data-based decision making, most specifically, the processes underlying teachers' ability to interpret or make sense of data. The data that teachers interpret in the current study are Curriculum-Based Measurement (CBM) data.

Curriculum-Based Measurement

CBM is a system that teachers use to monitor the progress of and evaluate the effectiveness of interventions for students with learning difficulties (Deno, 1985, 2003). CBM involves frequent, repeated, administration of short, simple measures that sample global performance in an academic area such as reading. CBM measures have been shown to be valid and reliable indicators of student performance and progress (see, for example, Wayman et al., 2007; Yeo, 2010; Shin and McMaster, 2019).

To assist teachers in interpreting the data, CBM scores are placed on a progress graph that depicts student growth over time in response to various iterations of an intervention (see Figure 1). The graph consists of: (a) *baseline data*, representing the student's beginning level of performance in comparison to peers; (b) a *long-range goal*, representing the desired level of performance at the end of the school year; (c) a *goal line* drawn from the baseline to the long-range goal, representing the desired rate of progress across the year; (d) *data points* representing the student's performance on weekly measurement probes; (e) *phases of instruction* separated by vertical lines, representing the initial intervention and adjustments to that intervention, and; (f) *slope lines*, representing the student's rate of growth within each instructional phase.

The progress graph lies at the heart of CBM because it guides teachers' instructional decision-making (Deno, 1985). When using CBM, teachers regularly inspect the CBM graph

to evaluate student progress within each phase of instruction. Based on their interpretation of the data, teachers make one of the following instructional decisions:

- (1) *Modify/adjust the intervention*, when the slope line is below and/or less steep than the goal line, indicating that the student is performing below the expected level and/or progressing at a rate slower than expected;
- (2) *Continue the intervention as is*, when the slope line is at a level equal to and parallel to the goal line, indicating that the student is progressing at the expected level and rate of progress;
- (3) *Raise the goal*, when the slope line is above and parallel to or steeper than the goal line, indicating that the student is progressing above the expected level and/or progressing more rapidly than expected.

Once the teacher has made an instructional decision, the teacher implements the decision, and then continues to collect data to evaluate the effects of the decision on student progress. This ongoing cycle of data interpretation, instructional decision-making, data interpretation, instructional decision-making, etc. is an integral part of Data-based Instruction (DBI; Deno, 1985; National Center on Intensive Intervention, 2013). When implemented appropriately, DBI results in individually tailored interventions for students with learning difficulties that, in turn, lead to significant improvements in the academic performance of the students (Filderman et al., 2018; Jung et al., 2018). Implementing DBI "appropriately," however, requires that teachers accurately read and interpret the CBM progress graphs.

Curriculum-Based Measurement graph comprehension

The ability to read and interpret—to "derive meaning from"—graphs is referred to as graph comprehension (Friel et al., 2001, p. 132). Graph comprehension can be influenced by both the characteristics of the graph and the viewer (Friel et al., 2001). Regarding the viewer—which is the focus of the present study—research has demonstrated that preservice and inservice teachers have difficulty describing CBM graphs in an accurate, complete, and coherent manner (Espin et al., 2017; van den Bosch et al., 2017; Wagner et al., 2017; Zeuch et al., 2017). For example, van den Bosch et al. (2017) found that inservice teachers were less complete and coherent in describing CBM graphs than CBM experts, and were less likely than the experts to compare student data to the goal line, to compare data across instructional phases, and to link data to instruction. Making such comparisons and links are essential for using CBM data to guide instruction.

Although research has made it clear that teachers have difficulty comprehending CBM graphs, it has not made clear

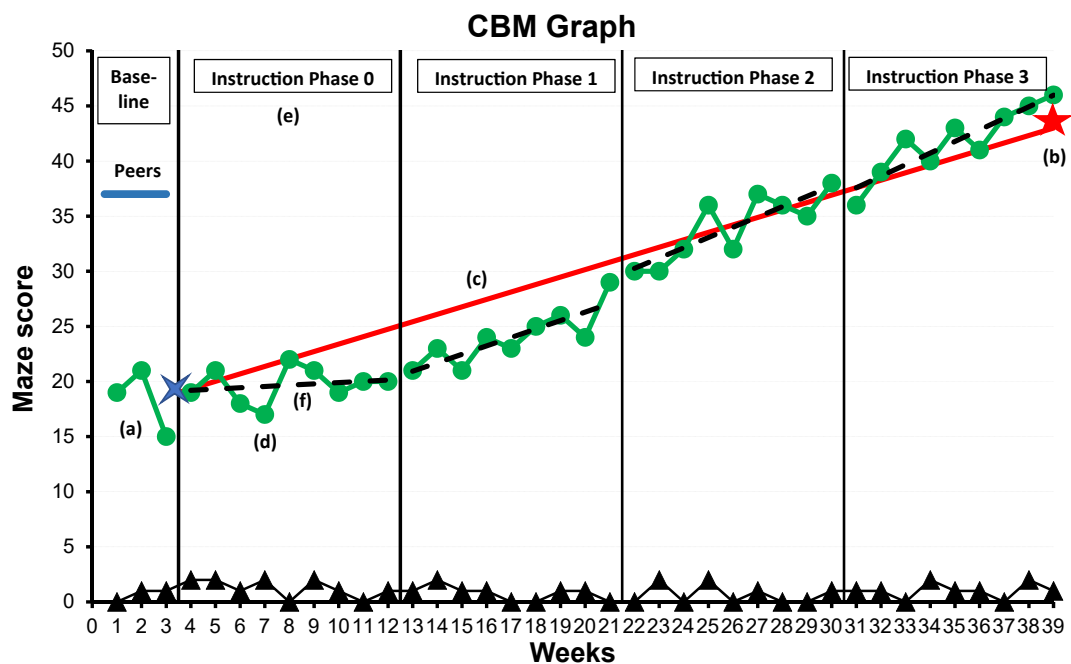


FIGURE 1

Sample CBM progress graph: (a) baseline data; (b) long-range goal; (c) goal line; (d) data points; (e) phase of instruction; (f) slope (growth) line.

why teachers have such difficulties. Little is known about the processes underlying teachers' ability to read and interpret CBM progress graphs. Knowing more about these processes might help to pinpoint where problems lie and might provide insights into how to improve teachers' graph comprehension. One technique for gaining insight into the processes underlying completion of visual tasks such as graph reading is eye-tracking.

Eye-tracking

Eye-tracking is a technology used to register people's eye movements while completing a visual task. Eye-movements reveal how attention is allocated when viewing a stimulus to complete a task and provide insight into the cognitive strategies used to complete the task (Duchowski, 2017). Eye-tracking has been used in reading to gain understanding of and insight into the processes underlying the reading of text (e.g., see Rayner, 1998; Rayner et al., 2006). Specific to teacher behaviors, eye-tracking has been used to study teachers' visual perception of classroom events (van den Bogert et al., 2014), awareness of student misbehavior (Yamamoto and Imai-Matsumura, 2012), and perceptions of problematic classroom situations (Wolff et al., 2016). In the area of graph reading, eye-tracking has been used to gain understanding into the processes underlying interpretation of graphs and to examine differences in processes related to the type and complexity of the graph (Vonder Embse, 1987; Carpenter and Shah, 1998; Okan et al., 2016).

The current study is to the best of our knowledge the first to use eye-tracking to study teachers' reading of CBM progress graphs. As such, it is an exploratory, descriptive study. Because there were no previous studies to guide us, the first challenge we faced in designing the study was to know what to expect of teachers. To address this challenge, we collected eye-tracking data from a member of the research team with expertise in CBM (see section "Materials and method") to provide a frame of reference for interpreting the teachers' data. A second challenge was to determine which variables to consider when analyzing the eye-tracking data. To address this challenge, we drew upon previous eye-tracking studies that compared experts' and novices' eye movements.

Eye-movements: Experts vs. novices

Across a wide variety of fields including medicine, sports, biology, meteorology, forensics, reading, and teaching, eye-tracking has been used to examine differences between experts and novices in their comprehension of visual stimuli and their use of strategies used to complete visual tasks (e.g., see Canham and Hegarty, 2010; Jarodzka et al., 2010; Al-Moteri et al., 2017; Watalingam et al., 2017; Beach and McConnel, 2019). A consistent finding to emerge from these studies is that experts devote more attention to task-relevant parts of visual stimuli and approach visual tasks in

a more goal-directed or systematic manner than do novices. Similar findings have emerged from eye-tracking research on the comprehension of graphs. For example, [Vonder Embse \(1987\)](#) found that experts fixated significantly longer on important parts of mathematical graphs than did novices, and that these differences were related to overall comprehension of the graphs. Similarly, [Okan et al. \(2016\)](#) found that viewers with high graph literacy devoted more time to viewing relevant features of graphs than participants with low graph literacy.

Drawing upon this previous body of eye-tracking research, we decided to examine the extent to which teachers devoted attention to various elements of CBM progress graphs and the extent to which they viewed the graphs in a systematic, orderly manner.

Purpose of the study

This study was an exploratory, descriptive study aimed at describing teachers' patterns of visual inspection when reading and interpreting CBM progress graphs. Teachers viewed CBM progress graphs and completed a think-aloud in which they described what they were looking at. As they completed their think-alouds, teachers' eye-movements were registered. Results from the think-aloud portion of the study have been reported elsewhere ([van den Bosch et al., 2017](#)). In this paper, we focus on the eye-tracking data. Our overall purpose is to develop and illustrate a method that can be used to examine teachers' inspection of CBM progress graphs and to delineate potential patterns of visual inspection that can be more closely examined in future research.

Our general research question was: *What are teachers' patterns of visual inspection when reading and interpreting CBM progress graphs?* We addressed three specific research questions:

1. To what extent do teachers devote attention to various elements of CBM graphs?
2. To what extent do teachers inspect the elements of CBM graphs in a logical, sequential manner?
3. Do the visual inspection patterns examined in research questions 1 and 2 differ for teachers with higher- vs. lower-quality graph descriptions (i.e., think-alouds)?

Materials and methods

Participants

Teachers

Participants were 17 fifth- and sixth-grade teachers (15 female; $M_{age} = 42.9$ years, $SD = 11.77$, range: 26–60) from

eight different schools in the Netherlands, who were recruited *via* convenience sampling. The original sample consisted of 19 teachers. Inspection of the demographic data collected from the teachers revealed that two of the teachers had completed a university course on CBM prior to the study. Because none of the other participating teachers had prior knowledge of or experience with CBM, we decided to exclude the data for these two teachers from the study.

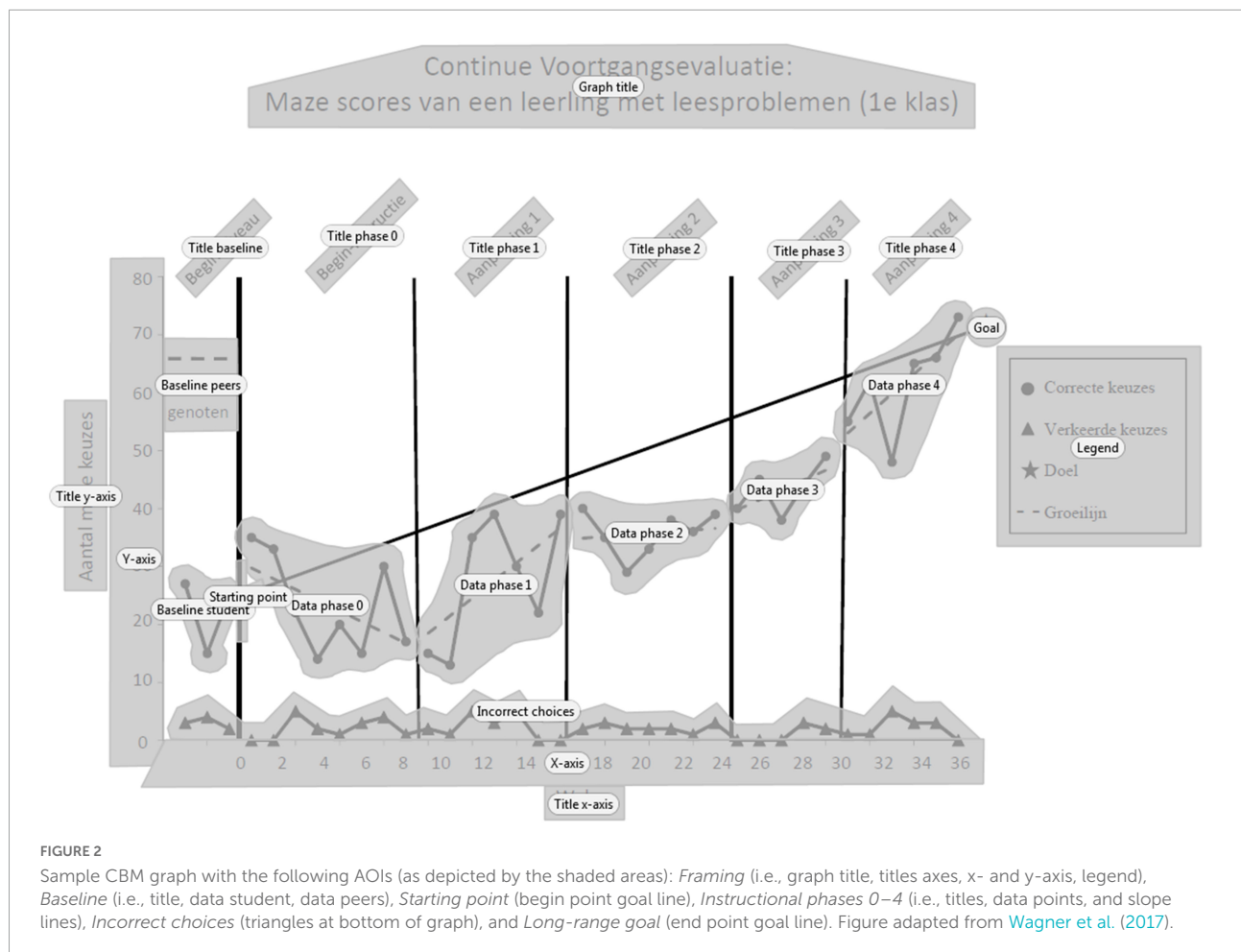
Participating teachers had all completed a teacher education program and held bachelor's degrees in education. One teacher also held a master's degree in psychology. Teachers had on average 17.82 years ($SD = 10.11$, range: 5–37) of teaching experience. All teachers had students with reading difficulties/dyslexia in their classes. Although the 17 participating teachers were not familiar with CBM prior to the start of the study, they were familiar with the general concept of progress monitoring because Dutch elementary-school teachers are required to monitor the progress of their students *via* standardized tests given one to two times per school year.

Curriculum-Based Measurement expert

To provide a frame of reference for interpreting the teachers' data, a member of the research team with expertise in CBM completed the same eye-tracking task as the teachers prior to the start of the study. The CBM expert was a university professor in the area of learning disabilities, with a Ph.D in educational psychology/special education, and with more than 23 years of experience conducting research and training on CBM, and with more than 40 publications focused on CBM and/or reading interventions for students with learning disabilities.

Materials: Curriculum-Based Measurement graphs

Two researcher-made CBM graphs were used in the study. The graphs depicted fictitious but realistic student data and were designed to capture data patterns often seen in CBM progress graphs. The data points and data patterns differed across the two graphs, but the set-up for each graph was the same, and included baseline data for the student and peers, a long-range goal, a goal line, five phases of instruction (labeled as Phases 0–4), data points, slope (growth) lines drawn through the data points within each phase, and a legend (see sample graph, [Figure 2](#); note that the graphs shown to the participants did not have any shaded areas). The order in which the two graphs were presented was counterbalanced (AB vs. BA) across teachers. The graphs for this study were modified versions of those used in [Wagner et al. \(2017\)](#). The graph titles, scales, and labels were changed to reflect CBM maze-selection rather than reading-aloud and were written in Dutch.



Eye-tracking procedures

To examine teachers' patterns of graph inspection, their eye-movements were registered as they described each graph. Prior to describing the graphs, teachers were shown a sample CBM graph and given a short description of the graph. They were told that the graph depicted the reading progress of one student receiving intensive reading instruction, and that the scores on the graph represented the student's correct and incorrect choices on weekly administered 2-min maze-selection probes. Each graph element was identified and described briefly to the teachers (see van den Bosch et al., 2017, for the full description).

Teachers were then positioned in front of the eye-tracker screen. They were told that they would be shown a CBM graph and that they would be asked to "think out loud" while looking at the graph. They were asked to tell all they were seeing and thinking, including what they were looking at and why they were looking at it. After calibrating the eye-tracker, instructions were repeated, and the first graph was presented. After teachers had described the first graph, the graph was removed from the screen, the instructions were again repeated,

and the other graph was presented. There were no time limits for the graph descriptions.

Data were collected in individual sessions at the teachers' schools by trained doctoral students and a trained research assistant. Two data collectors were present during each data collection session. One data collector operated the eye-tracker while the other instructed the participant and audio-taped the think-aloud. Instructions were read aloud from a script.

Eye-tracking apparatus and software

To register the eye movements of the participants, a Tobii T120 remote eye tracker was used. The Tobii T120 Eye Tracker is robust with regard to participants' head movements and its calibration procedure is quick and simple (Tobii Technology, 2010). Participants were positioned in front of the Tobii eye-tracker screen so that the distance between their eyes and the screen was approximately 60 cm. The data sampling rate was set at 60 Hz. The accuracy of the Tobii T120 Eye Tracker typically is 0.5 degrees, which implies an average error of 0.5

centimeter between the measured and the actual gaze direction (Tobii Technology, 2010).

Tobii Studio 3.4.8 and IBM SPSS Statistics 23 were used to process and to descriptively analyze the eye-tracking data.

Eye-tracking data

Establishing areas of interest

To analyze the eye-tracking data, Areas of Interest (AOIs) were defined for the graphs (see [Figure 2](#), shaded areas). We categorized the AOIs into 10 graph elements: (1) *Framing* (areas related to the graph set-up, including graph title, x- and y-axes and titles, the legend); (2) *Baseline* (areas related to baseline data, including title, baseline student, baseline peers); (3) *Starting point* (beginning point of the goal line), (4)–(8): *Instructional phases 0, 1, 2, 3, and 4*, respectively (areas related to instructional phases, including titles, data points, and slope lines within each phase), (9) *Incorrect choices* (triangles at the bottom of graph), and (10) *Long-range goal* (end point of the goal line). These ten graph elements were similar to those identified in previous research on CBM graph comprehension (Espin et al., 2017; Wagner et al., 2017), and to those coded in the think-aloud portion of the study (see van den Bosch et al., 2017). Due to the nature of CBM graphs in which different graph elements are near each other, some of the AOIs were adjacent to each other, and in some instances, overlapped slightly.

Fixation duration and fixation sequence

Two types of eye-tracking data were examined in this study: fixation duration data and fixation sequence data. *Fixations* serve as measures of visual attention and are defined as a short period of time in which the eyes remain still to perceive a stimulus, that is, to cognitively process the stimulus (Holmqvist et al., 2011). *Fixation duration* is the sum of the duration of all fixations within a particular area of the stimulus and *fixation sequence* is the order in which participants look at each area.

Fixation duration served as an indicator of participants' distribution of visual attention. The minimal fixation duration setting was set to 200 ms, meaning that a fixation was not registered unless the participant looked at a specific point for at least 200 ms. This cutoff point was chosen because typical values for fixations range from 200 to 300 ms (Holmqvist et al., 2011). For each participant the total duration of fixations (in sec.) was computed for each AOI via the eye-tracker software, after which the percentage of visual attention devoted to each of the 10 graph elements was calculated.

Fixation sequence served as an indicator of the extent to which teachers inspected CBM graph elements in a logical, sequential manner. Fixation sequences revealed the order in which participants viewed the CBM graph elements. For each participant, the sequence of fixations was computed via the eye-tracker software. The fixation sequence data were provided in

the form of strings of graph element names (e.g., *Baseline, Phase 1, Phase 2, Phase 1, etc.*).

Coding teachers' visual inspection of Curriculum-Based Measurement graphs

Attention devoted to elements of the graph

To address research question 1, to what extent teachers devoted attention to various elements of the CBM graph, for each teacher, the total duration of fixations (in sec.) was computed for each AOI via the eye-tracker software, after which the percentage of visual attention devoted to each graph element was calculated. Percentages were then totaled across teachers.

Sequence of visual inspection patterns

To address research question 2, to what extent teachers inspected CBM graph elements in a logical, sequential manner, the extent to which teachers' sequence of fixations followed a logical, sequential order were examined. As a first step, an "ideal sequence" was created based on the order in which CBM graphs would be used for instructional decision-making (see Espin et al., 2017; Wagner et al., 2017). The teachers' sequence of fixations was then compared to this ideal sequence. The ideal sequence used in the study was similar to the ideal sequence used to code the think-aloud data from the CBM graph descriptions (see van den Bosch et al., 2017). The ideal sequence for the eye-tracking data was: *Framing* (i.e., fixating on the elements related to the set-up of the graph), *Baseline*, *Goal setting*, *Instructional phases 0, 1, 2, 3, and 4*, and *Goal achievement*. The element *incorrect choices* was not included in the sequential analysis because it spanned multiple phases. Further, because participants could inspect the long-range goal either as a part of goal setting or goal achievement, the following rule was applied: If participants fixated on the long-range goal prior to fixating on any of the instructional phases, the fixation was coded under goal setting. If participants fixated on the long-range goal after fixating on at least one instructional phase, the fixation was coded as goal achievement.

The coding sheet presented in [Figure 3](#) was used to code the percentage of teachers' sequences following the ideal sequence. Along the top and down the left side of the coding sheet, the graph elements are listed. Sequences between graph elements were recorded using tally marks. To illustrate, let us assume that the viewer examined the graph elements in the following order: *Framing—Baseline—Goal setting—Baseline—Phase 0—Phase 1—Phase 2—Phase 1—Phase 3—Phase 4—Phase 3—Goal achievement*. The first viewing sequence in this example is Framing (FR) to Baseline (BL). This is recorded in the coding sheet in [Figure 3](#) with a tally mark at the intersection of the FR row and the BL column. The second sequence is Baseline (BL) to Goal setting (GS), which is recorded with

	FR	BL	GS	P0	P1	P2	P3	P4	GA
FR		I							
BL			I	I					
GS		I							
P0					I				
P1						I	I		
P2					I				
P3								I	I
P4							I		
GA									

FIGURE 3

Coding sheet for calculating the logical sequence percentages from the fixation sequence data. FR, Framing; BL, Baseline; GS, Goal setting; P0, Instructional phase 0; P1, Instructional phase 1; P2, Instructional phase 2; P3, Instructional phase 3; P4, Instructional phase 4; GA, Goal achievement. Figure adapted from [Espin et al. \(2017\)](#).

a tally mark at the intersection of the BL row and the GS column, and so forth. After all sequences were recorded on the coding sheet, the percentage of sequences following the ideal sequence was calculated.

Fixation sequences (strict approach)

The ideal sequence (*Framing to Baseline to Goal setting to Instructional phases 0–4, to Goal achievement*) is depicted by the light gray boxes above the diagonal in [Figure 3](#). To determine the percentage of sequences following the ideal sequence, the number of tallies in the light gray boxes was divided by the total number of tallies. The greater the percentage of tallies in the light gray boxes above the diagonal, the more closely the participant's graph inspection matched the ideal sequence. In the example in [Figure 3](#), five of 11 sequences fall in the light gray boxes above the diagonal, resulting in a logical sequence percentage of 45.5%. We refer to this approach as the “strict” calculation approach. After calculating the sequences using this strict approach, we calculated sequences using a more liberal approach that took into account lookbacks between adjacent graph elements.

Fixation sequence (liberal approach)

As we were coding the fixation sequences, we observed that the teachers (as well as the CBM expert) often looked back and

forth between adjacent graph elements—in particular, between adjacent instructional phases. Looking back and forth between graph elements (lookbacks) might reflect the fact that a viewer is comparing information across elements. Such comparisons are an important aspect of higher-level graph comprehension ([Friel et al., 2001](#)), and are an essential aspect of CBM data-based decision-making (see [van den Bosch et al., 2017](#)). We thus decided to calculate the fixation sequences in a second, more liberal, manner that took into account “lookbacks” between adjacent graph elements.

To calculate logical sequences using the liberal approach, we counted the number of tallies in the light gray boxes directly above and below the diagonal, and then divided this number by the total number of tallies. In the example in [Figure 3](#), eight of 11 sequences fell in the light gray boxes either above or below the diagonal, resulting in a logical sequence percentage of 72.7% for the liberal approach. We also counted the subset of tallies between adjacent *instructional phases* only (as opposed to between all graph elements). Comparing data across adjacent instructional phases (e.g., P1 to P2 or P2 to P1) is essential for determining whether instructional adjustments have been effective. In the example in [Figure 3](#), five of the 11 instances of lookbacks were between adjacent instructional phases, resulting in an instructional phase lookback percentage of 45.5%.

Intercoder agreement

The fixation sequence data for all participants were coded by a trained doctoral student and a trained research assistant. Intercoder agreement was 99.94%. There was one disagreement between coders, which was resolved through discussion.

Visual inspection patterns: Higher- vs. lower quality graph descriptions

Question 3 addressed whether the visual inspection patterns examined in research questions 1 and 2 differed for teachers with higher- vs. lower-quality graph descriptions. By addressing this question, we were able to link the eye-tracking data to the think-aloud data. Recall that teachers described the graphs *via* a think-aloud procedure while their eye-movements were being registered. These think-alouds were then compared to the think-alouds of three CBM experts (different from the expert in this study; see van den Bosch et al., 2017).¹ For the current study, we selected the two teachers with the highest-quality think-alouds (i.e., most similar to think-alouds of the experts), and the two teachers with the lowest-quality think-alouds (i.e., least similar to think-alouds of the experts) and compared their patterns of graph inspection.

Results

Fixation duration: Attention devoted to Curriculum-Based Measurement graph elements

The first research question was: To what extent do teachers devote attention to various elements of CBM

graphs? Data are reported as average scores across the two graphs. The overall viewing time for the teachers was on average 107.91 sec per graph ($SD = 59.83$; range 53–252.5 sec). This was compared to 283 sec per graph for the CBM expert.

The percentages of visual attention (i.e., fixation duration) devoted to each graph element for the teachers are reported in Table 1. Data for the CBM expert also are reported to provide a frame of reference (columns 2 and 1, respectively). Teachers devoted a fair amount of visual attention to FR (approximately 26%), which was similar to the value for the CBM expert (approximately 23%). Teachers devoted the largest proportion of visual attention to the five phases of instruction (approximately 58%), and, except for Phase 3, devoted approximately equal amounts of attention to each phase (approximately 11–14%). This pattern was somewhat different from that of the CBM expert. The expert also devoted the largest proportion of visual attention to the phases of instruction, but the percentage was larger than that of the teachers (approximately 70%). In addition, the expert did not devote equal amounts of attention to each phase, but rather devoted an increasing amount of attention across phases, devoting little attention to Phase 0 (approximately 3%), and much more attention to Phase 4 (approximately 25%). Finally, teachers devoted approximately 7.5 and 7% to Baseline and Incorrect choices, respectively, compared to 4 and 0.2% for the CBM expert.

Fixation sequence: Logical sequence of visual inspection patterns

The second research question of the study was: To what extent do teachers inspect the elements of the CBM graphs in a logical, sequential manner? Recall that we calculated fixation sequence using both a strict and liberal approach,

¹ Unfortunately, no eye-tracking data could be collected from the three experts involved in the think-aloud portion of the study. For this reason, we collected eye-tracking data from a different CBM expert for the current study.

TABLE 1 Mean percentage of visual attention devoted to graph elements for CBM expert, all teachers, and HQ-TA and LQ-TA teachers.

	CBM expert ($n = 1$)	All teachers ($n = 17$)	HQ-TA teachers ($n = 2$)	LQ-TA teachers ($n = 2$)
Graph elements (Areas of interest)				
Framing	22.66	25.61 (10.54)	17.27	27.08
Baseline	3.70	7.42 (4.31)	6.22	12.06
Starting point	0.56	0.46 (0.93)	0	0.20
Instructional phase 0	2.57	12.48 (6.81)	8.94	9.57
Instructional phase 1	13.23	13.81 (4.21)	15.97	9.19
Instructional phase 2	13.52	11.23 (3.82)	12.58	12.86
Instructional phase 3	15.68	6.99 (2.79)	10.88	3.19
Instructional phase 4	24.59	13.04 (6.27)	17.72	13.2
<i>TOTAL: Instructional phases 0–4</i>	69.59	57.55 (9.39)	66.09	48.02
Long-range goal	3.27	2.07 (2.16)	1.92	2.04
Incorrect choices	0.22	6.89 (4.52)	8.49	10.61

In the “All teachers column,” standard deviations are provided in parentheses.

HQ-TA teachers, teachers with higher-quality think-alouds; LQ-TA teachers, teachers with lower-quality think-alouds.

and also calculated the percentage of lookbacks between adjacent instructional phases only. Using the strict calculation approach, the mean logical sequence percentage for the teachers was 24.59% ($SD = 5.96$, range: 15.78–37.50). Using the liberal calculation approach, it was 40% ($SD = 10.8$, range: 18.16–56.49). These percentages were smaller than the 40.83 and 74.11% for the CBM expert for the strict and liberal approaches, respectively. The mean percentage of lookbacks between adjacent instructional phases for the teachers was 30.44% ($SD = 9.7$) compared to 49.62% for the CBM expert.

Visual inspection patterns: Higher- vs. lower quality graph descriptions

Visual attention data for the two teachers with higher- and lower-quality think alouds (HQ-TA and LQ-TA) are reported in the last two columns of [Table 1](#). The data reveal that HQ-TA teachers spent a smaller proportion of time viewing Framing and Baseline than did LQ-TA teachers (approximately 17 vs. 27%, respectively, for Framing, and 6 vs. 12%, respectively, for Baseline), and a larger proportion of time viewing the five instructional phases (approximately 66 vs. 48%, respectively).

We also compared the fixation sequence data for the teachers with higher- and lower- quality think-alouds. Using the strict calculation approach, the mean logical sequence percentage for HQ-TA teachers was 33.57%, compared to 16.94% for the LQ-TA teachers. Using the liberal calculation approach, the mean logical sequence percentage for HQ-TA teachers was 50.39%, compared to 30.19% for the LQ-TA teachers. Finally, the mean percentages of lookbacks between adjacent instructional phases for the HQ-TA was 37.75%, compared to 18.13% for the LQ-TA teachers.

Discussion

The purpose of this study was to examine how teachers visually inspected CBM graphs, and thereby, to gain insight into the processes underlying teachers CBM graph comprehension. The three research questions we addressed in the study were: (1) To what extent do teachers devote attention to various elements of CBM graphs? (2) To what extent do teachers inspect the elements of CBM graphs in a logical, sequential manner? (3) Do the visual inspection patterns examined in research questions 1 and 2 differ for teachers with higher- vs. lower-quality graph descriptions (i.e., think-alouds)? To provide a frame of reference for interpreting the teachers' data, data also were collected from a member of the research team who was a CBM expert.

Teachers' visual inspection of Curriculum-Based Measurement graph elements

The overall viewing time per graph for the teachers was about 2.5 times shorter than for the CBM expert. Given that the task had no time limits, the differences are notable, and suggest that teachers inspected the graphs in a less detailed manner than did the expert.

Examinations of the distribution of visual attention provide more insight into these differences. Teachers devoted most of their visual attention (58%) to the data in the instructional phases, which may not be that surprising given that 5 of the 10 graph elements that were categorized in AOIs were instructional phases. Nonetheless, it is positive that the teachers devoted a considerable amount of time to viewing the data in instructional phases. If teachers are to make sound data-based instructional decisions based on CBM graphed data, they must inspect the data within and between instructional phases to draw conclusions about student progress and the effectiveness of instruction. Despite this positive note, it is important to note the discrepancy between the teachers and CBM expert, who devoted nearly 70% of visual attention to the instructional phases, a much higher percentage than for the teachers. Further, there were differences between the teachers and CBM expert in the distribution of attention across the phases. Except for phase 3, teachers' attention was fairly evenly distributed across the phases, whereas the expert's attention increased across phases, from 3% in Phase 0 to 25% in Phase 4.

The discrepancies between the teachers and the CBM expert suggest that the teachers were less likely than the expert to focus attention on the most relevant aspects of the graph, a finding that fits with previous eye-tracking research. Previous research has shown that novices are less likely than experts to focus attention on relevant aspects of visual stimuli within the context of a task and are more likely to skim over non-relevant aspects (e.g., [Vonder Embse, 1987](#); [Canham and Hegarty, 2010](#); [Jarodzka et al., 2010](#); [Okan et al., 2016](#); [Al-Moteri et al., 2017](#)). With respect to the CBM graph, the most relevant areas of the graph are the *Instructional Phases* because they provide information on the effectiveness of instruction, and on the need to adjust that instruction. Within the instructional phases, the final phase is especially relevant because the data in this phase represent the overall success of the teacher's instruction across the school year, and signal whether the student will achieve the long-range goal.

Supporting the idea that teachers are less likely than the expert to focus on relevant aspects of the graph and more likely to focus on irrelevant aspects of the graph, is the percentages of visual attention devoted to *Incorrect choices*. Teachers focused nearly 7% of their visual attention on *Incorrect choices*, compared to 0% for the expert. Within CBM it is the

number of correct, not incorrect, choices that reflect growth. The number of incorrect choices is informational, but not as relevant for instructional decision-making as is the number of correct choices. Teachers may have tended to focus on incorrect choices because in typical classroom assessments, incorrect answers are used to calculate grades and to determine where students experience difficulties.

Logical sequence of visual inspection patterns

The second research question addressed the extent to which teachers inspected the CBM graphs elements in a logical (ideal) sequence; that is, a sequence that reflected the order in which CBM graphs would be used for instructional decision-making. For the teachers, 25% of their fixation sequences followed the ideal sequence, whereas for the CBM expert, it was 41%. Using a liberal calculation approach, which took into account looking back and forth between adjacent graph elements, the percentages were 40% for the teachers vs. 74% for the CBM expert. These results reveal that, as a group, teachers viewed the graphs in a less logical, sequential manner than the expert. The results are in line with previous eye-tracking studies comparing experts and novices that have shown that experts are more systematic and goal-directed in completing a visual task than novices (e.g., Jarodzka et al., 2010; Al-Motiri et al., 2017). The results also fit with the think-aloud data from the larger study, and with previous CBM graph comprehension research, which have shown that preservice and inservice teachers describe CBM graphs in a less logical, sequential manner than do CBM experts (van den Bosch et al., 2017; Wagner et al., 2017).

Differences between teachers and the CBM expert also were seen in the percentage of lookbacks between adjacent instructional phases. For teachers, 30% of their fixation sequences involved lookbacks between adjacent instructional phases, whereas for the CBM expert it was 50% of the sequences. These results suggest that the teachers did not often visually compare data points and slope lines between adjacent instructional phases, something that is important for making decisions about the effectiveness of instructional adjustments. These results again mirror the results of the think-aloud data, which showed that teachers were less likely to make data-to-data comparisons than were CBM experts in the larger study (van den Bosch et al., 2017).

In sum, the eye-tracking data indicate which aspects of CBM graph reading may be most problematic for the teachers and most in need of attention when teachers are learning to implement CBM. Specifically, the results suggest that teachers may need to learn to devote more attention to relevant aspects of the graphs such as the instructional phases (especially the later phases), and less attention to irrelevant aspects of the graphs, such as incorrect choices. Furthermore, teachers may need to

learn how to view graph elements in a sequence that reflects the time-sensitive nature of the graph. They may also need to learn to compare graph elements, especially how to compare data and slope lines across adjacent phases of instruction, so that they can use the data to evaluate the effects of instruction and of instructional adjustments.

Visual inspection patterns: High- vs. low-quality think-alouds

By comparing visual inspection patterns for teachers with higher- and lower-quality think alouds, we were able to link the eye-tracking data to teachers' ability to accurately and coherently describe CBM graphs. In general, the results demonstrated that visual inspection patterns for teachers with high-quality think alouds were more similar to those of the expert than visual inspection patterns for teachers with low-quality think alouds. Regarding fixation duration data, the HQ-TA teachers devoted more attention to the data in the instructional phases than did the LQ-TA teachers, with a difference of nearly 18%. With regard to the fixation sequence data, the HQ-TA teachers inspected the CBM graph elements in a more logical sequence, regardless of whether the strict or liberal calculation approach was used, and had a larger percentage of lookbacks between adjacent instructional phases, than did the LQ-TA teachers. Differences between the HQ-TA and LQ-TA teachers were approximately 20% for all three measures.

These data suggest that some teachers struggle more than others in reading, interpreting, and comprehending CBM graphs. The think-aloud data for the LQ-TA teachers (see van den Bosch et al., 2017) had shown that they were not able to describe CBM graphs in a complete and coherent manner and did not make within-data comparisons when describing the graphs. These differences were reflected in these teachers' patterns of graph inspection. The LQ-TA teachers spent relatively little time on the most relevant aspects of the graph (i.e., instructional phases), and inspected the graphs in a less logical, sequential manner than did the HQ-TA teachers. Further, based on the lookback data, the LQ-TA teachers did not appear to make comparisons between adjacent phases of instruction. In short, although results of this study suggest that all teachers might benefit from specific, directed instruction in reading and comprehending CBM progress graphs, the data comparing the HQ-TA and LQ-TA teachers suggest that some teachers will need more intensive and directed instruction than others.

Limitations

The present study was an exploratory, descriptive study that used eye-tracking technology to examine teachers' patterns of

inspection when reading CBM graphs. Results of this study should be viewed as a springboard for developing future studies with larger and more diverse samples. The study had several limitations. First, the sample was a small sample of convenience, and consisted of teachers with relatively little experience with CBM. Although appropriate for an exploratory study, it is important to replicate the study with a larger more representative sample, and with teachers who have used CBM for an extended period of time. Second, the data used to provide a frame of reference were collected from only one CBM expert, and this expert was a member of the research team who was familiar with the graphs used in the study. Although the graph descriptions of this expert were nearly identical to the graph descriptions given by the three CBM experts from the think-aloud portion of the study [who were not familiar with the graphs (van den Bosch et al., 2017)] it is still a limitation. The study should be replicated with other CBM experts.

Third, the AOIs were in some cases adjacent to each other or even overlapped slightly. We elected to use graphs that were set up identically to those used in the Wagner et al. (2017) so that we could tie our data to that earlier study. These graphs had ecological validity in that they were typical of the type of progress graphs actually seen by teachers when using CBM. That said, bordering/overlapping AOIs are not desirable in analyzing eye-tracking data, and thus the data patterns found in this explorative study should be viewed as suggestive, and should be verified in future research with graphs that are designed so that AOIs do not border on/overlap with each other.

Implications for practice and for future research

Although teachers are expected to closely monitor the progress of students with severe and persistent learning difficulties, and to evaluate the effectiveness of the given instruction for these students with systems like CBM, the results of the present study suggest that teachers have difficulty inspecting CBM graphs, with some teachers having more difficulty than others. Combining the results of the current study with the results of previous think-aloud studies on CBM graph reading (Espin et al., 2017; van den Bosch et al., 2017; Wagner et al., 2017), the results suggest the need to provide teachers with specific, directed instruction on how to inspect, read, and interpret CBM graphs. Unfortunately, such instruction may not typically be a part of CBM professional development training (see Espin et al., 2021b), which is worrisome given that student achievement improves only when teachers adequately respond to CBM data with instructional and goal changes (see Stecker et al., 2005).

Graph-reading instruction could be improved in different ways. For example, teachers could be taught where to direct their attention when reading CBM graphs.

Keller and Junghans (2017) used such an approach for helping viewers to read medical graphs and demonstrated that providing the viewers with written instructions on reading medical graphs while arrows pointed to the task-relevant parts of the graphs increased visual attention for the task-relevant graph parts. Alternatively, teachers could be shown a video of the eye-movements of a CBM expert completing a think-aloud description of a CBM graph. The video would illustrate how to inspect the graph in a detailed, logical, sequential manner. Such Eye Movement Modeling Examples (EMMEs) have been used in other areas such as medical education (Jarodzka et al., 2012; Seppänen and Gegenfurtner, 2012) and digital reading (Salmerón and Llorens, 2019). Teachers' ability to read and interpret progress graphs could also be improved *via* specific, directed instruction focused on CBM graph reading, combined with multiple practice opportunities, as demonstrated by van den Bosch et al. (2019).

A final method of improving teachers' ability to read and interpret CBM progress graphs would be to design the graphs in a way to direct teacher attention to key elements of the graph and to provide graph-reading supports. For example, the slopes could be presented in different colors that correspond to the decision to be made (red for adjusting instruction, green for keeping instruction as is, blue for raising the goal), or graph elements could be hidden or highlighted with a click of the mouse. For a review of graph supports that have been effective in assisting teachers in CBM decision-making (see Stecker et al., 2005; Fuchs et al., 2021).

Conclusion

The results of this exploratory, descriptive study provide insights into how teachers visually inspect CBM progress-monitoring graphs and provide a basis for designing future studies focused on teachers' ability to read and interpret student progress graphs. The results of the present study revealed differences between teachers and the CBM expert, and between teachers with higher- and lower-quality think-alouds, in terms of how long participants inspected relevant graph elements and the order in which they inspected the elements. In comparison to the expert, teachers as a group were found to be less adept at focusing on the relevant aspects of the CBM graphs, at inspecting the graph elements in a logical sequence, and at comparing data across adjacent instructional phases. However, there were differences between teachers: Teachers in this study who produced better descriptions of the CBM graphs (HQ-TA teachers) were more adept at graph inspection than teachers who produced poorer descriptions of the graphs (LQ-TA teachers). The results of this study, in combination with the results of think-aloud studies of CBM graph comprehension, highlight potential areas of need for teachers, and provide guidance regarding the design of CBM instruction for teachers.

Before making firm conclusions about teachers' inspection of CBM progress graphs, it will be important to replicate the present study with a larger and more diverse sample, and with independent CBM experts. An important aspect of future research will be to tie teachers' graph descriptions and patterns of graph inspection to their actual use of CBM data for instructional decision-making, and, ultimately, to student achievement. With a larger data set, statistical models could also be applied to the data (see for example, [Man and Harring, 2021](#)) to determine whether particular processing patterns/profiles predict teachers' CBM graph comprehension, teachers' appropriate use the data for instructional decision making, and, ultimately, student achievement.

Data availability statement

The data supporting the conclusions of this article will be made available by the authors for specific purposes upon request.

Ethics statement

This study was reviewed and approved by the Ethics Committee, Education and Child Studies. Written informed consent to participate in the study was provided by the participants.

References

- Al-Moteri, M. O., Symmons, M., Plummer, V., and Cooper, S. (2017). Eye tracking to investigate cue processing in medical decision-making: A scoping review. *Comput. Hum. Behav.* 66, 52–66. doi: 10.1016/j.chb.2016.09.022
- Beach, P., and McConnel, J. (2019). Eye tracking methodology for studying teacher learning: A review of the research. *Int. J. Res. Method Edu.* 42, 485–501. doi: 10.1080/1743727X.2018.1496415
- Beck, J. S., and Nunnaley, D. (2021). A continuum of data literacy for teaching. *Stud. Educ. Eval.* 69:100871.
- Canham, M., and Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learn. Instr.* 20, 155–166. doi: 10.1016/j.learninstruc.2009.02.014
- Carpenter, P. A., and Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *J. Exp. Psychol. Appl.* 4, 75–100. doi: 10.1037/1076-898X.4.2.75
- Datnow, A., and Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *J. Educ. Change* 17, 7–28. doi: 10.1007/s10833-015-9264-2
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Except. Child.* 52, 219–232. doi: 10.1177/001440298505200303
- Deno, S. L. (2003). Developments in curriculum-based measurement. *J. Spec. Educ.* 37, 184–192. doi: 10.1177/00224669030370030801
- Deno, S. L. (2013). "Problem-solving assessment," in *Assessment for Intervention: A Problem-Solving Approach*, 2nd Edn, eds R. Brown-Chidsey and K. J. Andren (New York, NY: Guilford Press), 10–36.
- Duchowski, A. T. (2017). *Eye Tracking Methodology: Theory and Practice*, 3rd Edn. Cham: Springer International Publishing.
- Espin, C. A., Förster, N., and Mol, S. (2021a). International perspectives on understanding and improving teachers' data-based instruction and decision making: Introduction to the special series. *J. Learn. Disabil.* 54, 239–242. doi: 10.1177/00222194211017531
- Espin, C. A., van den Bosch, R. M., van der Liende, M., Rippe, R. C. A., Beutick, M., Langa, A., et al. (2021b). A systematic review of CBM professional development materials: Are teachers receiving sufficient instruction in data-based decision-making?. *J. Learn. Disabil.* 54, 256–268. doi: 10.1177/0022219421997103
- Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., and de Rooij, M. (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learn. Disabil. Res. Pract.* 32, 8–21. doi: 10.1111/ldrp.12123
- Filderman, M. J., Toste, J. R., Didion, L. A., Peng, P., and Clemens, N. H. (2018). Data-based decision making in reading interventions: A synthesis and meta-analysis of the effects for struggling readers. *J. Spec. Educ.* 52, 174–187. doi: 10.1177/0022466918790001

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

Acknowledgments

We would like to thank the teachers who participated in this research, also Anouk Bakker for her help with the preparation of this manuscript, and Arnout Koornneef for feedback on the design and analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Friel, S. N., Curcio, F. R., and Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *J. Res. Math. Educ.* 32, 124–158. doi: 10.2307/749671
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Stecker, P. (2021). Bringing data-based individualization to scale: A call for the next-generation of teacher supports. *J. Learn. Disabil.* 54, 319–333. doi: 10.1177/0022219420950654
- Gleason, P., Crissey, S., Chojnacki, G., Zukiewicz, M., Silva, T., Costelloe, S., et al. (2019). *Evaluation of Support for Using Student Data to Inform Teachers' Instruction*. (NCEE 2019-4008). Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. New York, NY: Oxford University Press.
- Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., et al. (2012). Conveying clinical reasoning based on visual observation via eye-movement modelling examples. *Instr. Sci.* 40, 813–827. doi: 10.1007/s11251-012-9218-5
- Jarodzka, H., Scheiter, K., Gerjets, P., and van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learn. Instr.* 20, 146–154. doi: 10.1016/j.learninstruc.2009.02.019
- Jung, P.-G., McMaster, K. L., and delMas, R. C. (2017). Effects of early writing intervention delivered within a data-based instruction framework. *Except. Child.* 83, 281–297. doi: 10.1177/0014402916667586
- Jung, P.-G., McMaster, K. L., Kunkel, A. K., Shin, J., and Stecker, P. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-analysis. *Learn. Disabil. Res. Pract.* 33, 144–155. doi: 10.1111/ldrp.12172
- Keller, C., and Junghans, A. (2017). Does guiding toward task-relevant information help improve graph processing and graph comprehension of individuals with low or high numeracy? An eye-tracker experiment. *Med. Decis. Making* 37, 942–954. doi: 10.1177/0272989X17713437
- Kuchle, L. B., Edmonds, R. Z., Danielson, L. C., Peterson, A., and Riley-Tillman, T. C. (2015). The next big idea: A framework for integrated academic and behavioral intensive intervention. *Learn. Disabil. Res. Pract.* 30, 150–158. doi: 10.1111/ldrp.12084
- Man, K., and Haring, J. R. (2021). Assessing preknowledge cheating via innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educ. Psychol. Meas.* 81, 441–465.
- Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educ. Psychol.* 47, 71–85. doi: 10.1080/00461520.2012.667064
- Mandinach, E. B., and Schildkamp, K. (2021). Misconceptions about data-based decision making in education: An exploration of the literature. *Stud. Educ. Eval.* 69:100842. doi: 10.1016/j.stueduc.2020.100842
- National Center on Intensive Intervention (2013). *Data-Based Individualization: A Framework for Intensive Intervention*. Washington, DC: Office of Special Education, U.S. Department of Education.
- Okan, Y., Galesic, M., and García-Retamero, R. (2016). How people with low and high graph literacy process health graphs: Evidence from eye-tracking. *J. Behav. Decis. Making* 29, 271–294. doi: 10.1002/bdm.1891
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037/0033-2909.124.3.372
- Rayner, K., Chace, K. H., Slattery, T. J., and Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Sci. Stud. Read.* 10, 241–255. doi: 10.1207/s1532799xssr1003_3
- Salmerón, L., and Llorens, A. (2019). Instruction of digital reading strategies based on eye-movements modeling examples. *J. Educ. Comput. Res.* 57, 343–359. doi: 10.1177/0735633117751605
- Schildkamp, K., Ehren, M., and Lai, M. K. (2012). Editorial article for the special issue on data-based decision making around the world: From policy to practice to results. *Sch. Eff. Sch. Improv.* 23, 123–131. doi: 10.1080/09243453.2011.652122
- Seppänen, M., and Gegenfurtner, A. (2012). Seeing through a teacher's eye improves students' imagining interpretation. *Med. Educ.* 46, 1113–1114. doi: 10.1111/medu.12041
- Shin, J., and McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *J. Sch. Psychol.* 73, 131–149. doi: 10.1016/j.jsp.2019.03.005
- Stecker, P. M., Fuchs, L. S., and Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychol. Sch.* 42, 795–819. doi: 10.1002/pits.20113
- Tobii Technology (2010). *Tobii T/X Series Eye Trackers: Product Description*. Available online at: <https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/tobii-pro-tx-product-description.pdf>
- van den Bogert, N., van Bruggen, J., Kostons, D., and Jochems, W. (2014). First steps into understanding teachers' visual perception of classroom events. *Teach. Teach. Educ.* 37, 208–216. doi: 10.1016/j.tate.2013.09.001
- van den Bosch, R., Espin, C. A., Chung, S., and Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learn. Disabil. Res. Pract.* 32, 46–60. doi: 10.1111/ldrp.12122
- van den Bosch, R. M., Espin, C. A., Pat-El, R. J., and Saab, N. (2019). Improving teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *J. Learn. Disabil.* 52, 413–427. doi: 10.1177/0022219419856013
- Vanlommel, K., Van Gasse, R., Vanhoof, J., and Van Petegem, P. (2021). Sorting pupils into their next educational track: How strongly do teachers rely on data-based or intuitive processes when they make the transition decision?. *Stud. Educ. Eval.* 69:100865.
- Vonder Embse, C. B. (1987). *An Eye Fixation Study of Time Factors Comparing Experts and Novices When Reading and Interpreting Mathematical Graphs*. Ph.D. thesis. Columbus, OH: Ohio State University.
- Wagner, D. L., Hammerschmidt-Snidarich, S., Espin, C. A., Seifert, K., and McMaster, K. L. (2017). Pre-service teachers' interpretation of CBM progress monitoring data. *Learn. Disabil. Res. Pract.* 32, 22–31. doi: 10.1111/ldrp.12125
- Watalingam, R. D., Richetelli, N., Pelz, J. B., and Speir, J. A. (2017). Eye tracking to evaluate evidence recognition in crime scene investigations. *Forensic Sci. Int.* 208, 64–80. doi: 10.1016/j.forsciint.2017.08.012
- Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., and Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *J. Spec. Educ.* 41, 85–120. doi: 10.1177/00224669070410020401
- Wolff, C. E., Jarodzka, H., van den Bogert, N., and Boshuizen, H. P. A. (2016). Teacher vision: Expert and novice teachers' perception of problematic classroom management scenes. *Instr. Sci.* 44, 243–265. doi: 10.1177/0022487114549810
- Yamamoto, T., and Imai-Matsumura, K. (2012). Teachers' gaze and awareness of students' behavior: Using an eye tracker. *Compr. Psychol.* 2:6. doi: 10.2466/01.IT.6
- Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial Spec. Educ.* 31, 412–422. doi: 10.1177/0741932508327463
- Zeuch, N., Förster, N., and Souvignier, E. (2017). Assessing teachers' comprehension to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learn. Disabil. Res. Pract.* 32, 61–70. doi: 10.1111/ldrp.12126



OPEN ACCESS

EDITED BY
Stefan Blumenthal,
University of Rostock, Germany

REVIEWED BY
Jürgen Wilbert,
University of Potsdam, Germany
Jeffrey M. DeVries,
Technical University Dortmund,
Germany

*CORRESPONDENCE
Boris Forthmann
boris.forthmann@wwwu.de

SPECIALTY SECTION
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Education

RECEIVED 14 April 2022
ACCEPTED 28 September 2022
PUBLISHED 03 November 2022

CITATION
Forthmann B, Förster N and
Souvignier E (2022) Multilevel
and empirical reliability estimates
of learning growth: A simulation study
and empirical illustration.
Front. Educ. 7:920704.
doi: 10.3389/feduc.2022.920704

COPYRIGHT
© 2022 Forthmann, Förster and
Souvignier. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Multilevel and empirical reliability estimates of learning growth: A simulation study and empirical illustration

Boris Forthmann*, Natalie Förster and Elmar Souvignier

Institute of Psychology in Education, University of Münster, Münster, Germany

Reliable learning progress information is crucial for teachers' interpretation and data-based decision making in everyday classrooms. Slope estimates obtained from simple regression modeling or more complex latent growth models are typically used in this context as indicators of learning progress. Research on progress monitoring has used mainly two ways to estimate reliability of learning progress, namely (a) split-half reliability and (b) multilevel reliability. In this work we introduce empirical reliability as another attractive alternative to quantify measurement precision of slope estimates (and intercepts) in learning progress monitoring research. Specifically, we extended previous work on slope reliability in two ways: (a) We evaluated in a simulation study how well multilevel reliability and empirical reliability work as estimates of slope reliability, and (b) we wanted to better understand reliability of slopes as a latent variable (by means of empirical reliability) vs. slopes as an observed variable (by means of multilevel reliability). Our simulation study demonstrates that reliability estimation works well over a variety of different simulation conditions, while at the same time conditions were identified in which reliability estimation was biased (i.e., with very poor data quality, eight measurement points, and when empirical reliability was estimated). Furthermore, we employ multilevel reliability and empirical reliability to estimate reliability of intercepts (i.e., initial level) and slopes for the quop-L2 test. Multilevel and empirical reliability estimates were comparable in size with only slight advantages for latent variable scores. Future avenues for research and practice are discussed.

KEYWORDS

progress monitoring, slope, growth, reliability, simulation, formative assessment

Introduction

Evaluation of student learning is a crucial component to inform progress monitoring (Silbergltt and Hintze, 2007). Progress monitoring has been connected in particular with curriculum-based measurement (CBM) as a well-known formative assessment approach in special education (Deno, 1985, 1987). However, progress monitoring

approaches for the entire classroom such as learning progress assessment exist (Souvignier et al., 2021). The overarching goal of progress monitoring is learning growth which is commonly assessed in the literature by estimating the linear slope (Silbergliitt and Hintze, 2007) across multiple assessment points (e.g., weekly measurements in CBM). Given the importance of these growth estimates for progress monitoring it is clear that their reliability needs to be as high as possible to most accurately inform teachers' instructional decisions.

Hence, issues related to the reliability of progress monitoring slopes such as schedule and duration (i.e., number of occasions per week and overall number of weeks of data collection), or dataset quality (as operationalized by the amount of residual variance in growth models) have been extensively examined in simulation studies (Christ et al., 2012, 2013a; Van Norman et al., 2013). One major dependent variable in such simulation studies is the true reliability of slope estimates (i.e., the squared correlation between estimated slopes and their true values). These studies have shown that acceptable levels of slope reliability (i.e., 0.70) can only be achieved for data collection durations of at least 6 or 8 weeks (depending further on the schedules; e.g., Christ et al., 2013a). A conclusion that was later backed-up with empirical data (Thornblad and Christ, 2014). For empirical data, however, the true slope values are not known and reliability of slopes can only be estimated. Yet, little is known on how well reliability estimation methods quantify true reliability. This question has not been in the focus of previous work on the reliability of progress monitoring slopes and here we seek to address this gap in the literature.

Furthermore, Van Norman and Parker (2018), for example, compared two commonly used methods to estimate slope reliability, namely split-half reliability and multilevel reliability. Both methods aim at quantifying reliability of slopes as an observed variable (i.e., not as latent variable). Slopes as a latent variable, however, can be obtained by means of empirical Bayes estimates, for example, and one might think that these latent variable estimates are more reliable as compared to slope estimates as an observed variable. Hence, we extend the set of used reliability estimation methods by examining empirical reliability which quantifies reliability of progress as a latent variable. Empirical reliability is borrowed from the item-response theory literature (Green et al., 1984; Brown and Croudace, 2015) and shares with multilevel reliability the feature that it is easy to calculate. In fact, multilevel reliability and empirical reliability can be estimated even in case that only few measurement points are available which prevents estimation based on the split-half method (e.g., for only three measurement points). Thus, the aim of our study was twofold: (a) we wanted to know how well reliability estimates actually quantify true reliability, and (b) we wanted to know how reliability estimated for slopes as a latent variable performs in comparison to reliability estimated for slopes as an observed variable.

Reliability of growth in progress monitoring

Progress monitoring requires multiple measurement points over time. Hence, factors that undermine comparisons of test results across time potentially undermine reliability of progress monitoring estimates. For example, Van Norman and Parker (2018) outline lack of measurement invariance (i.e., parallel test forms should display equal difficulty), characteristics of the data collection procedure (e.g., used instructions, changing test administrators, or varying testing environments), and the testing schedule (i.e., number of measurement points within a given period of time) as potentially influencing factors. To study these influencing factors and their potential link with growth reliability, reliability must be estimated. Yet, the statistical methods used to estimate growth reliability can also be a source of heterogeneity in reliability findings (Van Norman and Parker, 2018). The focus in previous work (see above), however, was on the method of growth estimation (e.g., differences in true reliability between various slope estimators; Bulut and Cormier, 2018) rather than the estimation of growth reliability (i.e., which method of estimating reliability best quantifies true reliability). Hence, this work seeks to address this gap in the literature.

Methods of assessing reliability of slopes

Perhaps most often researchers use the split-half odd-even method to estimate the reliability of student growth estimates (VanDerHeyden and Burns, 2008; Christ et al., 2013b; Van Norman et al., 2013). This method requires measurement timepoints to be splitted into the odd and even timepoints. Learning growth is then estimated separately by ordinary least squares regression, for example, for each set of timepoints and each student. Analogous to classical test theory in which reliability is conceptualized as test-test correlation (e.g., Haertel, 2006), split-half reliability is obtained from the correlation between slopes based on the odd measurement points (e.g., measurement points 1, 3, and 5) and the slopes based on the even measurement points (e.g., measurement points 2, 4, and 6).

Among other outcomes, previous simulation studies typically focus on true reliability as well as estimated split-half reliability (Christ et al., 2012, 2013b) and, thus, split-half reliability is the only method for which we know how well it works. The match between estimated split-half reliability and true reliability decreased as a function of number of measurement timepoints as well as data quality (operationalized as the amount of residual variance). Presumably, conditions with few measurement points or large residual variance are more likely to yield violations of the assumptions underlying split-half reliability, namely equal true-score and error variances between the test-halves (e.g., Haertel, 2006). However, while split-half reliability is among the recommended methods for the evaluation of slope reliability (National Center on Intensive Intervention, 2014), we do

not focus on the method in this work as it requires at least six measurement timepoints which limits its range of application.

Another method relies on the ratio of true slope variability and overall variability of (OLS) slopes (e.g., Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). This method has been also referred to as multilevel reliability (e.g., Schatschneider et al., 2008; Van Norman and Parker, 2018). Multilevel reliability tends to go to one when the number of measurement points is large (relating to collection duration and schedules) or in case strong inter-individual differences in learning progress exist (e.g., Raudenbush and Bryk, 2002). Van Norman and Parker used a random-intercept-random-slope model (e.g., Snijders and Bosker, 2012) to estimate between-student learning growth variance (i.e., true slope variability) and the variance of OLS slopes obtained for each child (i.e., observed slope variability). They found that multilevel reliability was larger than uncorrected split-half reliability for all examined levels of duration. Yet, given that uncorrected split-half refers to reliability of slopes based on only half the timepoints, this is not surprising. True reliability of OLS slopes has also been quantified in simulation studies on learning growth in the context of curriculum-based measurement as the squared correlation between estimated and true learning growth (Christ et al., 2012, 2013b). However, these studies did not estimate multilevel reliability. Thus, simulation studies on learning progress estimation have thus far not looked at how well multilevel reliability works as an estimate of the reliability of OLS slope. We address this gap in the current work.

Finally, it should be noted that the estimate of learning progress as a latent variable is used for estimation of multilevel reliability (National Center on Intensive Intervention, 2014; Van Norman and Parker, 2018), yet latent variable scores can also be obtained from random-intercept-random-slope models. For example, the R package lme4 (Bates et al., 2015)—which is often used in the progress monitoring literature (e.g., Parker et al., 2011; McMaster et al., 2017; Van Norman and Parker, 2018)—provides values for the unknown unobserved latent variable by means of conditional modes given the observed data and estimated other parameter values (Bates et al., 2015). Reliability of such latent variable scores (i.e., the squared correlation between the estimated scores and the true scores) can be estimated by marginal or empirical reliability (Green et al., 1984; Brown and Croudace, 2015). Empirical reliability is widely used in item-response theory applications (e.g., Forthmann et al., 2020b,c; Beisemann, 2022), for example.

Aim of the current work

The reliability of learning progress estimates (i.e., slopes) is critically important for progress monitoring assessment. The known and used methods to quantify slope reliability in the field

of progress monitoring may not be applicable to all contexts. For example, split-half reliability cannot be used when only three measurement points are available. Furthermore, multilevel reliability as the ratio of the estimated slope variance across students (i.e., an estimate of “true” variance) to the OLS slope variance provides an estimate of OLS slope reliability. OLS slopes, however, are not always the best choice (e.g., when outliers are present; Bulut and Cormier, 2018). When data at hand require more complex modeling choices with respect to progress monitoring, empirical reliability might be another reasonable choice for slope reliability estimation. Empirical reliability—like multilevel reliability—can be used with at least three measurement points and can be understood as an estimate of the squared correlation between slope estimates and their unknown true values. In other words, it provides an estimate of slopes as a latent variable. Hence, increasing the awareness of researchers in the field that this approach to estimate reliability is available and provides useful psychometric information is the main aim of our paper. In accordance with this aim, we sought to complement existing simulation studies in the field of progress monitoring by examining how well multilevel reliability estimation (as an established method for learning progress reliability estimation; National Center on Intensive Intervention, 2014) and empirical reliability estimation work in a range of conditions used in previous simulation studies (Christ et al., 2012, 2013a; Van Norman et al., 2013). Finally, for illustration purposes, we apply the reliability estimation to real data using the quop-L2 test which is used in the context of learning progress assessment in everyday school contexts (Souvignier et al., 2021).

Simulation study

Simulation design

The simulation design is adapted from Christ et al. (2012) to connect with previous simulation studies. The design was based on the factors sample size with four levels ($N = 125$, $N = 250$, $N = 500$, and $N = 1,000$), data quality with the levels *very poor* and *very good* (referring to residual variances of $\sigma_e^2 = 25$ and $\sigma_e^2 = 400$, respectively), and number of timepoints ($T = 8$ and $T = 20$). Simulations were based on the following latent growth model:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij} + \epsilon_{ij} \quad (1)$$

with Y_{ij} being the test performance of child i ($i = 1, \dots, N$) at timepoint j ($j = 1, \dots, T$), latent variable means β_0 (i.e., the average intercept) and β_1 (i.e., the average slope), latent variable values b_{0i} (i.e., a child's deviation from the average intercept) and b_{1i} (i.e., a child's deviation from the average slope), and residual term ϵ_{ij} . Latent variables were bivariate normal with $\mu = (\beta_0, \beta_1)$ and covariance matrix $\Sigma = \begin{pmatrix} \sigma_{b_0}^2 & \sigma_{b_0 b_1} \\ \sigma_{b_0 b_1} & \sigma_{b_1}^2 \end{pmatrix}$. Average

intercept and average slope were set to $\beta_0 = 40$ and $\beta_1 = 1.5$ with variances $\sigma_{b_0}^2 = 150$ and $\sigma_{b_1}^2 = 0.40$, respectively. The correlation between intercept and slope was set to 0.20 for all simulations. Simulations were run by means of the R package *simsem* (Pornprasertmanit et al., 2020). We ran 1,000 replications for each cell of the simulation design. The R code is openly available in the Open Science Framework¹.

Dependent variables

We analyzed the following dependent variables:

- True reliability: The squared correlation between the true latent variables and their estimated values (i.e., either latent or observed).
- Estimated reliability: The estimated reliability by either empirical or multilevel reliability estimates.
- Bias: The difference between estimated and true reliability.
- RMSE (root mean squared error): The square-root of the squared difference between estimated and true reliability divided by the number of replications.

We mainly display simulation results graphically.

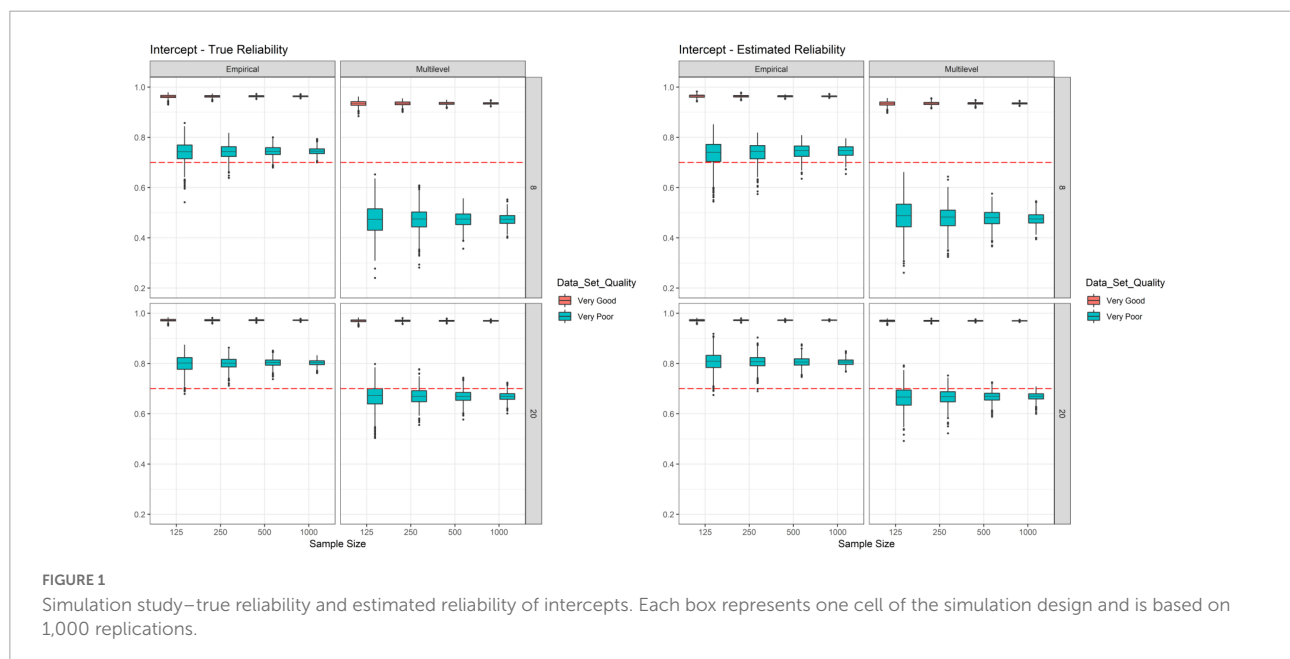
Results and discussion

Figure 1 displays true (left side) and estimated (right side) intercept reliability. Overall, true reliability for intercepts was

substantially stronger for very good data quality as compared to very bad data quality. We also found that true intercept reliability increased with the number of measurement points, yet this effect was clearly better visible for very poor data quality compared to very good data quality, and for multilevel reliability compared to empirical reliability. Sample size further decreased the variability of true intercept reliability. Again, this effect was clearly better visible for very poor data quality compared to very good data quality and for multilevel reliability compared to empirical reliability. As expected, the difference between empirical and multilevel reliability decreased as a function of data quality and number of measurement points. For example, for very good data quality and 20 measurement points reliabilities were clearly on par (see bottom-left in Figure 1), yet when looking at poor data quality and eight measurement points empirical reliability (i.e., the squared correlation between latent variable estimates and the true values) was substantially higher as compared to multilevel reliability (i.e., the squared correlation between OLS estimates and the true values; see top-left in Figure 1). Finally, it should be noted that with respect to true reliability we found that intercept reliability was below 0.70 only for very poor data quality and when multilevel reliability was estimated. The right side in Figure 1 demonstrated that estimated intercept reliability worked quite well. Indeed, estimated reliability pretty much mimicked the findings for true reliability pointing toward unbiased estimation of intercept reliability.

However, for true slope reliability (see left side in Figure 2) we found that reliabilities were only higher than 0.70 for very good data quality and when 20 measurement points were used. The average true multilevel reliabilities replicated the findings of Christ et al. (2012) well. For example, for

¹ <https://osf.io/mn5hx>



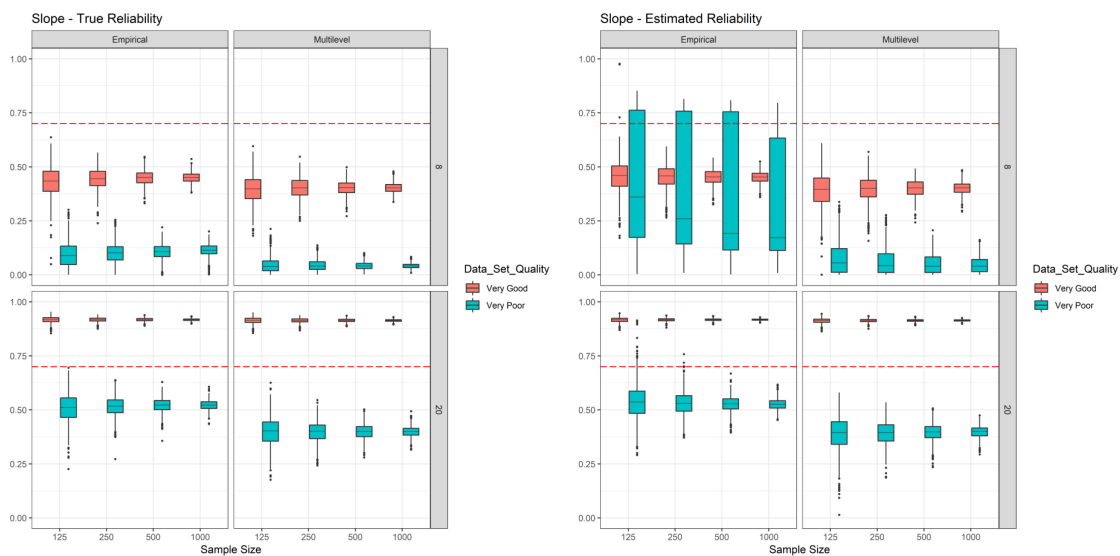


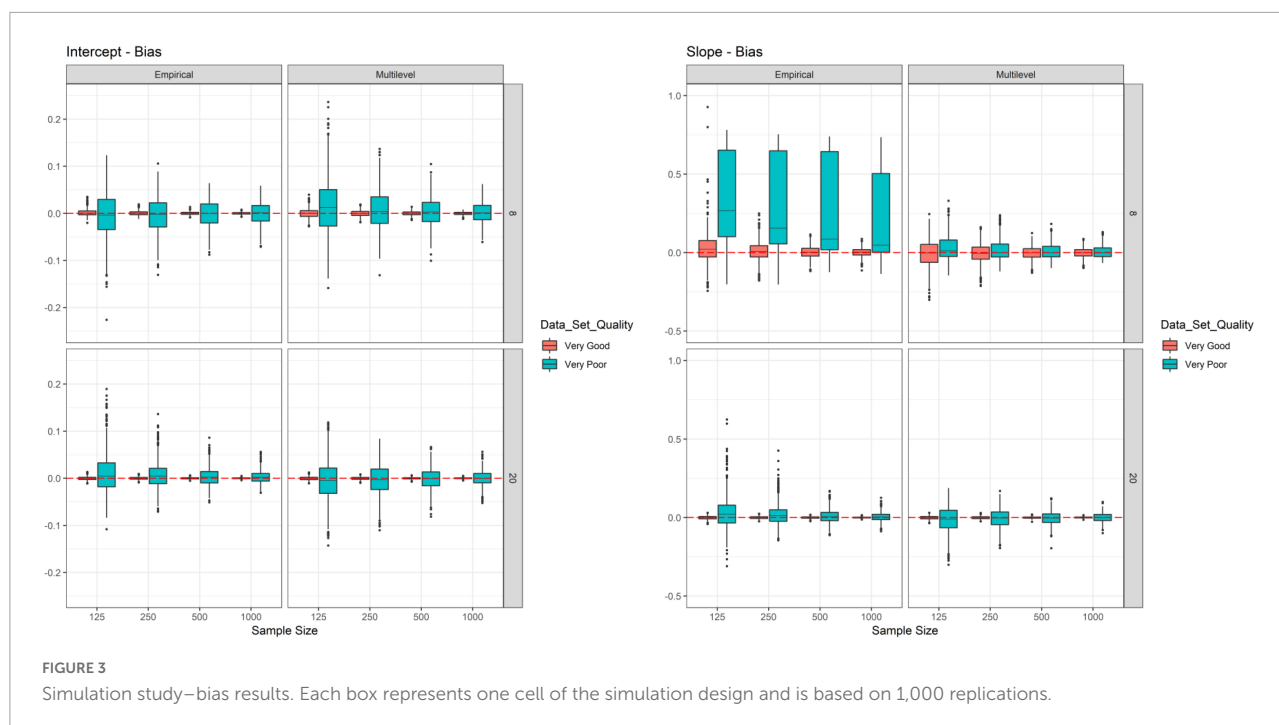
FIGURE 2

Simulation study—true reliability and estimated reliability of slopes. Each box represents one cell of the simulation design and is based on 1,000 replications.

very poor data quality and eight measurement points we found an OLS slope reliability of 0.40 for all sample size conditions, Christ et al. found 0.38 (the slight difference can be explained by their much smaller simulated sample size), whereas for very good data quality and 20 measurement points we found a reliability of 0.91 (across all simulated sample sizes) and Christ et al. also reported 0.91. These observations emphasize that our simulation setup is well linked to previous simulation studies. In addition, as for intercept reliability a clear main effect of data quality was observed (see red vs. cyan colored boxes on the left side in Figure 2). We further observed a clear main effect of measurement timepoints. The difference between empirical and multilevel reliability was not as strong for slope reliability as compared to intercept reliability (still the difference was stronger for very poor data quality vs. very good data quality, but also for 8 measurement points vs. 20 measurement points). Yet, as expected, again empirical reliability tended to be higher than multilevel reliability. Similarly, sample size had an effect on variability of true slope reliabilities. Estimated slope reliabilities, however, did not follow the true slope reliability findings and thus differed to intercept reliability findings above. Especially empirical reliability with very poor data quality and eight measurement points was heavily positively biased, i.e., true reliability was found to be strongly overestimated. Differences between true and estimated slope reliability were not as extreme for multilevel reliability. These observations are further illustrated in Figure 3 which depicts the bias of the estimates. There are several other conditions associated with very poor data quality and the smallest sample size in which reliability tended to be overestimated (also for multilevel reliability

and intercept reliability; see Figure 3). Thus, under certain conditions empirical reliability will provide a far too optimistic estimation of slope reliability, whereas multilevel reliability will provide a conservative estimate. Other biases tended to be negligible.

Finally, we evaluated RMSE as another measure of reliability estimation accuracy (see Figure 4). It should be noted that RMSEs for intercepts and slopes cannot be directly compared because both are per design on a different scale. RMSE was again a function of data quality with smaller values resulting for very good data quality (vs. very poor data quality). The only exception from this observation was for slope multilevel reliability with eight measurement points. Here, the differences were only negligible small and the amount of the difference depended on sample size (ranging from no difference for $N = 125$ to the highest difference for $N = 1,000$). This pattern can be explained by the known outlier sensitivity of the RMSE as a measure of accuracy and the findings obtained for true and estimated slope multilevel reliability as shown in Figure 2. For example, estimated reliability for the sample size of $N = 125$ had much more extreme points at the lower tail of the distribution when data quality was very good (red-colored box), whereas much more extreme points at the upper tail of the distribution were observed for very poor data quality (cyan-colored box). These extreme values at the respective tails of the distributions of estimated slope multilevel reliabilities surpassed the respective tails of the distributions of true reliabilities. Overall, this pattern resulted in highly similar RMSEs. This pattern diminished with increasing sample sizes, but was still clearly observable for $N = 250$ and $N = 500$.



Empirical illustration

Materials and methods

Participants

The sample used in this work comprised of $N = 4,970$ second-grade school students (nested in 298 classes) taken from the 2018 cohort (i.e., school year 2018/2019) which were assessed by the quop-L2 test series (Förster and Kuhn, 2021; Förster et al., 2021). The students in the final sample had a mean age of 7.95 years ($SD = 0.48$), 53% were boys and 47% were girls, and 81% did not have a migration background whereas 19% had a migration background. Notably, the cohort included initially 6,000 students, yet 1,030 were excluded for various reasons (students from international schools: $n = 140$; students from a different grade level who were assigned to quop-L2: $n = 227$; students with an age below 6 years: $n = 3$; students with an age above 12: $n = 94$; students with missing values on all measurement points: $n = 333$; and duplicate cases: $n = 233$). The same sample has been used in a recent study with a different focus (Forthmann et al., 2022).

The quop-L2 test series for progress monitoring in reading

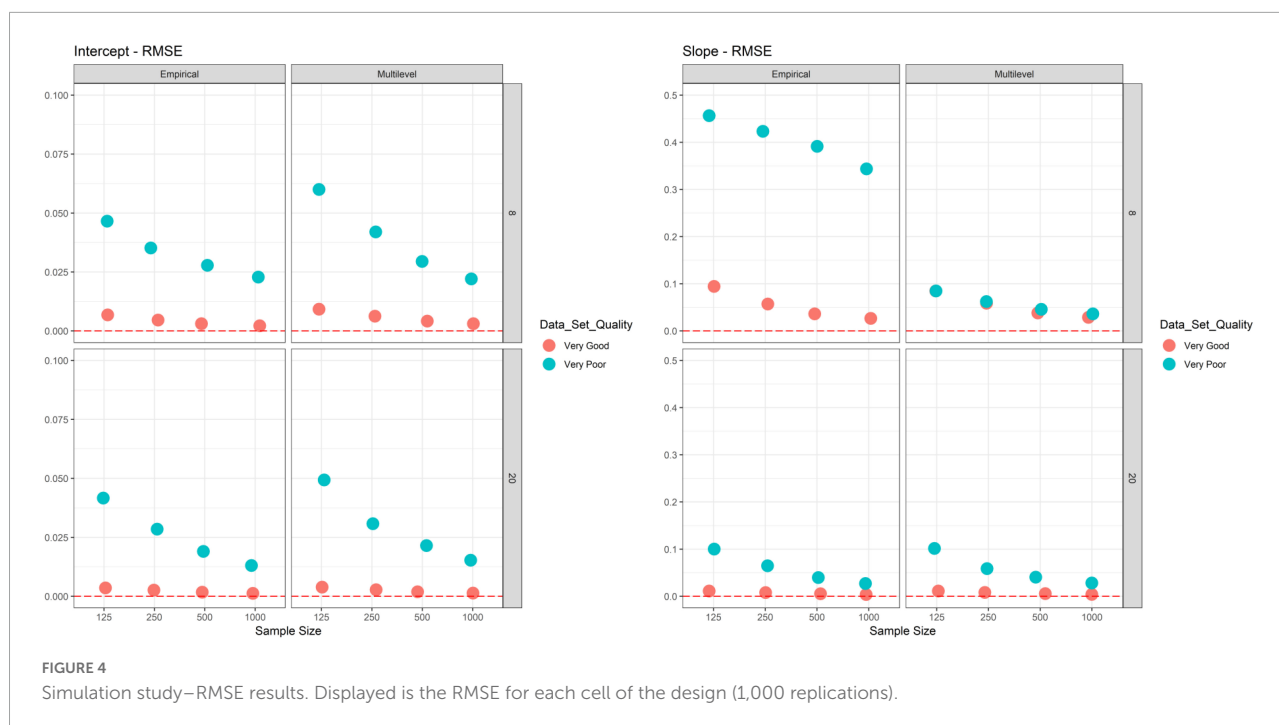
The quop-L2 test for reading achievement is comprised of four short equivalent versions with subscales at the word, sentence, and text level. The items of the tests were constructed based on three dichotomous item-features that determine item difficulty to a great extent. At the word level items were

word/pseudoword discrimination tasks (item features were number of syllables, word frequency, and the number of orthographic neighbors), sentence level items were sentence verification tasks (item features were propositional density, associations between target words, and complexity of the sentence structure), and items at the text level required a decision if a third sentence fits a story based on two initially presented sentences (item features were use of personal pronouns, content, and the presence of causal relationships). Each of the four tests included 20 word level items, 13 sentence-level items, and 13 text-level items. Each test was administered two times throughout the school year (i.e., there were eight measurement points). Students were randomly assigned to groups which received different combinations of test halves to prevent confounding of items and measurement points (Klein Entink et al., 2009). The eight measurement points of quop-L2 assessments were administered *via* the computerized quop assessment system (Souvignier et al., 2021). The tests were completed when students were studying on their own or in group sessions throughout the schoolyear. The quop-L2 tests displayed acceptable to excellent psychometric properties (Förster et al., 2021).

Analytical approach

All data and the analysis script to reproduce the reported findings in this work are openly available *via* a repository in the Open Science Framework².

² <https://osf.io/mn5hx>



To correct for fast guessing (Wise and DeMars, 2010; Wise, 2017) and unacceptable slow responding we used subscale specific quantiles as cut-offs for valid response behavior (fast guessing: 5%-quantile; slow responding: 99.5%-quantile). We obtained these quantiles across all items of each of the respective subscales (word level: lower bound = 1362.98 ms, upper bound = 41032.86 ms; sentence level: lower bound = 1427.02 ms, upper bound = 53742.18 ms; text level: lower bound = 877.36 ms, upper bound = 85836.71 ms). Item accuracy was scored after taking these cut-offs into account. The CISRT efficiency scoring was used to reflect reading achievement beyond accuracy (Maris and van der Maas, 2012). CISRT scoring requires item timing, but here assessment was untimed. Hence, the time cut-offs were used for CISRT scoring. Item scores were averaged for each subscale (i.e., word, sentence, and text level) and scaled to be in the range from 0 to 10.

The quop-L2 test series allows to model reading achievement as a higher-order factor based on word, sentence, and text level scores as observed indicators (Forthmann et al., 2022). Such an approach was also employed in the current work for the evaluation of longitudinally strong measurement invariance (Vandenberg and Lance, 2000) prior to growth modeling which is recommended in the progress monitoring literature (Schurig et al., 2021). This way comparisons across timepoints are not confounded by psychometric properties. Specifically, we evaluated three levels of measurement invariance: (a) configural invariance, (b) weak invariance, and (c) strong invariance. First, a configural model

was evaluated. Reading achievement was modeled as a latent variable at each of the eight measurement points by the three observed scores at word, sentence, and text level. For model identification purposes the loading of the sentence level score was fixed to one. Residuals of the observed scores were not allowed to covary, but all latent variable latent covariances were freely estimated. Next, this configural model (Model 1) was compared to two alternative configural models. In an alternative configural model (Model 2) we allowed residuals of the scores at the same level of language to covary (e.g., the residuals of all sentence scores were allowed to covary). However, the modeling of residual covariances for sentence level scores was empirically not supported and, hence, another configural model with residual covariances only at the word and text levels were considered (Model 3). This final model on which measurement invariance testing was based is depicted in Figure 5.

The statistical software R was used for data analysis (R Core Team, 2021). We used the lavaan package (Rosseel, 2012) for measurement invariance testing. Robust full information maximum likelihood estimation was employed for two reasons: (a) multivariate normality was violated, and (b) missing values were present in the data. Model fit was evaluated based on common cut-offs in the literature (West et al., 2012). Evidence in favor of strong measurement invariance for efficiency was already reported in detail by Forthmann et al. (2022) and we do not repeat the statistics here. Consequently, reading achievement as modeled in this work based on quop-L2 displayed time-invariant loadings and intercepts of observed indicators.

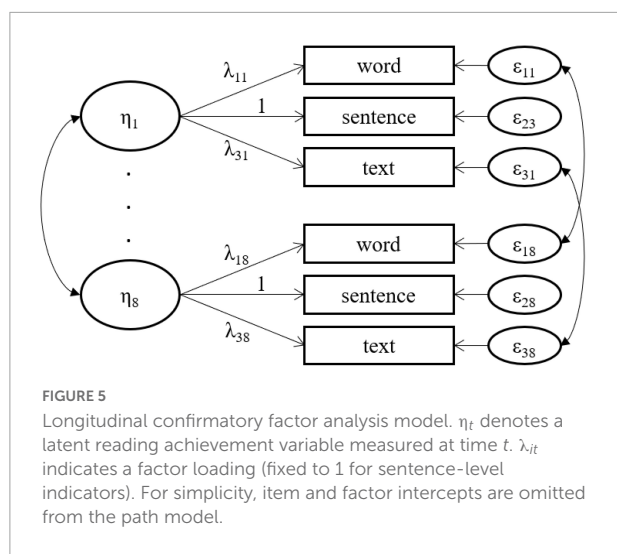


TABLE 1 Descriptive statistics and reliability estimates for reading achievement at each measurement timepoint.

Timepoint	<i>M</i>	<i>SD</i>	FDI	α	ω_1
T1	−0.60	1.36	0.85	0.76	0.77
T2	−0.33	1.29	0.86	0.76	0.77
T3	−0.08	1.31	0.87	0.78	0.78
T4	0.07	1.23	0.85	0.75	0.75
T5	0.26	1.23	0.84	0.75	0.75
T6	0.28	1.22	0.83	0.74	0.74
T7	0.42	1.20	0.85	0.76	0.76
T8	0.48	1.12	0.84	0.75	0.75

FDI = factor determinacy index. α = Cronbach's alpha. ω_1 = Bollen's estimate of congeneric composite reliability.

The Bartlett-method (DiStefano and Zhu, 2009) was used to estimate factor scores based on the longitudinally strong invariance models (i.e., one set of factor scores for each scoring). Factor determinacy indices (FDI) (Ferrando and Lorenzo-Seva, 2018), Cronbach's α (Cronbach, 1951), and Bollen's ω_1 (Bollen, 1980; Raykov, 2001) were further estimated and are reported in Table 1. The latter two coefficients were estimated by means of the semTools package (Jorgensen et al., 2021). Reliability of efficiency scores at each timepoint was larger than 0.70 as a recommended cut-off for low-stakes decisions (Christ et al., 2005) and all FDIs were greater than 0.80 see Ferrando and Lorenzo-Seva (2018).

In a next step, we subjected the factor scores to linear latent growth modeling (i.e., a random-intercept-random-slope model) which was estimated by means of the lme4 package (Bates et al., 2015). Intercept and slope varied across students. The timepoint variable in the analyses was coded in a way that allows to interpret the intercept as the initial level of reading achievement in the schoolyear (i.e., the first timepoint was coded as zero). Multilevel reliability was calculated as the

ratio of estimated slope variance to observed variance (i.e., the variance of individual OLS slope estimates; Van Norman and Parker, 2018). Finally, we obtained the slope variance from the estimated growth models and the average squared standard error of learning progress estimates for estimating empirical reliability (Brown and Croudace, 2015): Empirical Reliability = $1 - \hat{\sigma}_{b_{1i}, \text{Error}}^2 / \hat{\sigma}_{b_1}^2$. For completeness, we also assess reliability of the initial level estimates.

Results and discussion

Efficiency scores increased on each subsequent measurement point (see Table 1), while the standard deviation decreased over time. Figure 6 provides a graphical illustration of individual learning progress and the average growth which was slightly non-linear.

Initial level and slope reliability findings

The estimates of the random-intercept-random-slope model revealed an average intercept of −0.46 and an average slope of 0.15. Intercept and slope variances are reported in Table 2, with much higher intercept variation across students as compared to slope variation. The latent variable correlation between initial level and learning progress was found to be $r = -0.55$. Table 2 summarizes the initial level and slope reliability estimates. Reliability of intercept estimates was generally good to excellent, whereas slope reliability was comparably lower and below proposed cut-offs (e.g., 0.70). Then, as expected, it was further observed that multilevel reliability estimates were smaller as compared to empirical reliability. Yet, the observed differences were not large. In other words, latent variable scores were not much more reliable than observed OLS estimates.

To further check the trustworthiness of these reliability estimates, we reran the simulation based on the parameters obtained for the quop-L2 scores. As in the simulation study reported above, we ran 1,000 replications (the file to run this simulation is also available in the OSF repository). True multilevel reliability for intercept (0.85) and slope (0.41), as well as true empirical reliability for intercept (0.90) and slope (0.44) were highly comparable with the estimates obtained for the empirical data. In addition, estimated reliability for the simulated data matched true reliability very well. This was the case for intercept (0.85) and slope (0.41) multilevel reliability, as well as intercept (0.90) and slope (0.41) empirical reliability estimates. Thus, for the parameter estimates and the sample size of the quop-L2 data in this work, reliability estimation can be considered unbiased. In addition, the fact that reliability estimates obtained for the empirical data matched the simulation well further corroborates the impression of accurate reliability estimation for these data.

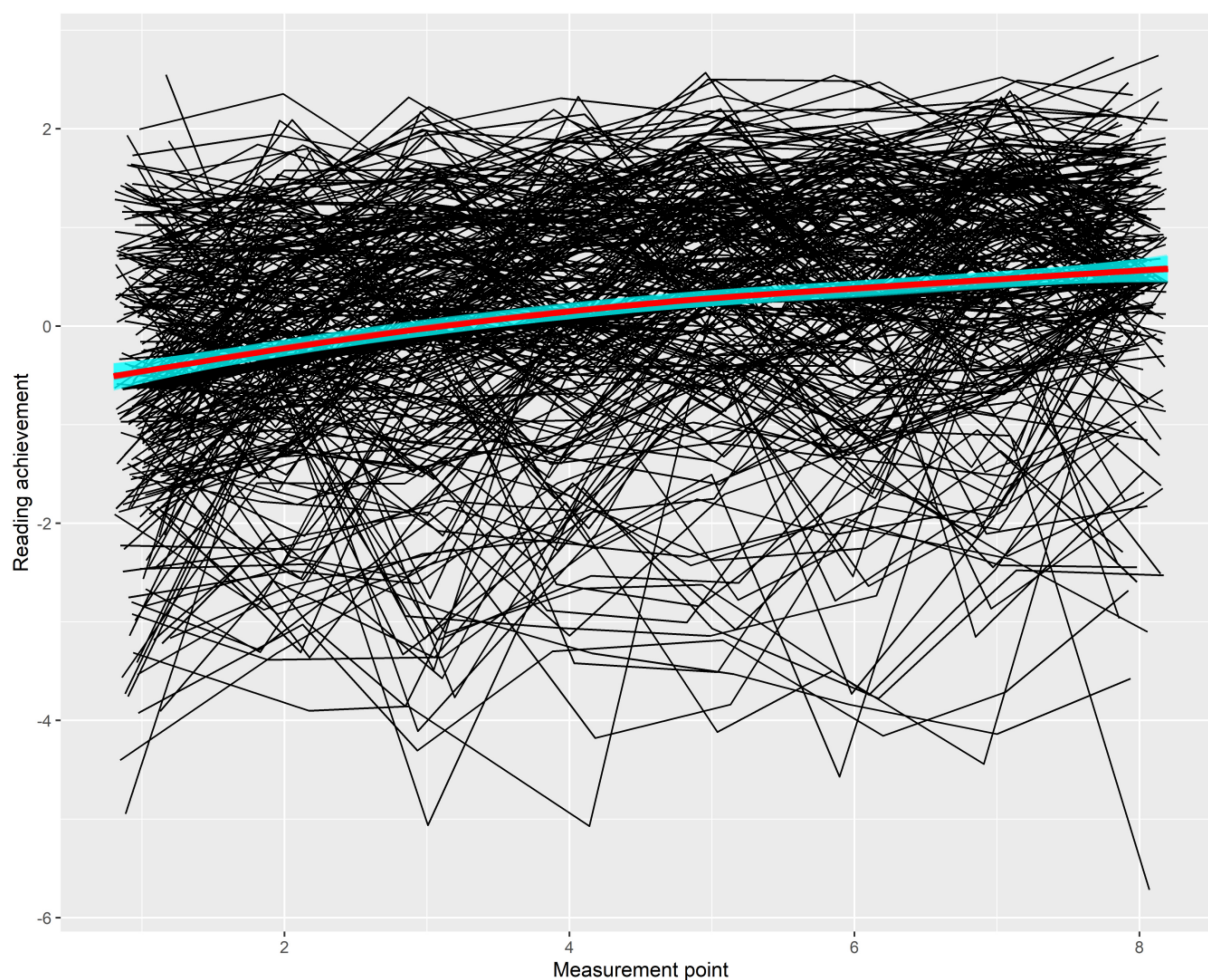


FIGURE 6

Spaghetti plot for learning growth trajectory of students. We chose trajectories of $n = 300$ students at random to increase the interpretability of the plot. Red line represents LOESS-smoothed average growth trajectory. Standard error band is shown in cyan.

Discussion

In this study, we examined how well reliability estimates actually quantify true reliability in a simulation study, and we evaluated more closely how reliability estimated for slopes as a latent variable performs in comparison to reliability estimated for slopes as an observed variable. Our simulation study revealed that estimation of multilevel as well as empirical reliability works well across a variety of conditions. Yet, especially conditions affected by very poor data quality, small sample size (i.e., $N = 125$), and/or rather few measurement points (i.e., eight measurement points) were found to result in slightly biased reliability estimation. In particular, empirical reliability estimates of learning progress was found to be upwardly biased when dataset quality was very poor and when only eight measurement points were available. Increasing sample size under such conditions did not remedy the observed bias.

We recommend that researchers use the openly available R scripts that come along with this paper to run a simulation based on obtained parameters for a given dataset. This should be especially done when data are found to be similar to the conditions in which reliability estimation was biased in our study. Overall, however, we conclude that reliability estimation works across a variety of simulation conditions used in previous work (Christ et al., 2012).

In addition, we estimated multilevel and empirical reliability for the quop-L2 reading test series which allows for progress monitoring in everyday classrooms (Förster et al., 2021; Souvignier et al., 2021). We found that multilevel and empirical reliability findings were similar in size to true and simulated reliability for eight measurement points and very good data quality in our simulation study. Relatedly, previous work estimated true multilevel reliability in simulation studies on slope estimation methods in the progress monitoring literature (Christ et al., 2012, 2013b; Christ and Desjardins, 2018). The

TABLE 2 Reliability estimates at the student level and at the class level.

	Intercept	Slope
$\hat{\sigma}_{b_0}^2$	1.293	–
$\hat{\sigma}_{b_1}^2$	–	0.009
$\hat{\sigma}_{b_{0,OLS}}^2$	1.515	–
$\hat{\sigma}_{b_{1,OLS}}^2$	–	0.021
$\hat{\sigma}_{Error}^2$	0.130	0.005
Multilevel reliability	0.853	0.406
Empirical reliability	0.900	0.438

Multilevel reliability for intercept = $\hat{\sigma}_{b_0}^2 / \hat{\sigma}_{b_{0,OLS}}^2$. Multilevel reliability for slope = $\hat{\sigma}_{b_1}^2 / \hat{\sigma}_{b_{1,OLS}}^2$. Empirical reliability for intercept = $1 - \hat{\sigma}_{b_{0i,Error}}^2 / \hat{\sigma}_{b_0}^2$. Empirical reliability for slope = $1 - \hat{\sigma}_{b_{1i,Error}}^2 / \hat{\sigma}_{b_1}^2$. All estimates are rounded to three decimals. Hence, not all reliability coefficients can be exactly calculated based on the reported estimates of the various variances because of rounding errors.

squared correlation between estimated slopes and their true values (i.e., reliability) has been commonly used as dependent variable in these simulation studies which is conceptually the same quantity that one is trying to estimate by multilevel reliability. These findings serve further as a benchmark for interpretation of the current findings. For example, researchers found a range for simulated *good* quality data and 8 weeks time schedule of 0.10 to 0.45 (Christ et al., 2012, 2013b). In light of these previous results one can again conclude that the findings in this study with a multilevel reliability of 0.41 again imply that reliability findings for quop-L2 provides are in accordance with reliability findings for progress monitoring data of good to very good quality.

Limitations and future directions

The main aim of this research was to extend previous work on the reliability of learning progress estimates by evaluating how well multilevel and empirical reliability work. Notably, empirical reliability as a way to quantify measurement precision has wide potential for applications in progress monitoring beyond the used simulation model and data used for illustration in this work. Yet, concrete findings reported here are limited to the conditions of our simulations and data which represent learning progress assessment as a form of progress monitoring in everyday school contexts. While this limitation is important when it comes to interpretations of the empirical findings in this work, we do not see that application of the approach in other forms of progress monitoring is undermined. Empirical reliability can readily be calculated as long as individual progress estimates and associated standard errors are available (e.g., when latent growth modeling is used).

We have discussed above findings from simulation studies on slope estimation approaches in the CBM literature. These findings might serve as a benchmark for the findings in this

work. In a sense, partially replicating previous work emphasizes their validity. However, it should not be overlooked that these simulations—and, hence, also the simulation study reported in this work—specify a set of population parameters for simulation that is informed by CBM research and not by learning progress assessment research. However, with this work, we provide open material that facilitates data simulations of progress monitoring data. Hence, we recommend running new simulations for other learning progress assessment conditions to complement interpretation of reliability estimates. Such a step is illustrated in this work and can be understood as a check of model fit. If simulated true and estimated reliabilities are far off the estimates obtained for a data set a cautious interpretation of findings is needed.

Conclusion

In this work we extended previous simulation studies on the reliability of learning progress assessment. First, previous work focused mainly on true reliability, whereas here we focused on how well reliability estimation works. Second, we additionally focused on empirical reliability as a way to quantify measurement precision of latent variable scores obtained from latent growth modeling. Overall, we found that reliability estimation works for a variety of conditions and recommend to check this locally by adapting our openly available simulation material. In addition, empirical vs. multilevel results may provide critical information to decide which estimate should be used in research and practice. For example, when OLS estimates turn out to be unreliable, latent variable estimates of learning progress might still be a useful option. For future work we recommend to estimate both types of reliabilities to be maximally informed.

Data availability statement

The datasets presented in this study can be found in online repositories. The name of the repository and accession number can be found below: Open Science Framework; <https://osf.io/mn5hx/>.

Ethics statement

This study was carried out in accordance with the recommendations by the Ethics Committee of the Department of Psychology of the University in Münster. Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. For participants involved in this study, either informed consent was obtained from their parents or their

participation was regulated based on a contractual regulation that allowed us to use participant data in an anonymized form for scientific purposes.

Author contributions

BF had the idea for the research reported in this work, analyzed the data, and wrote a first draft. NF was responsible for the development and research design of the quop-L2 test series. NF and ES contributed in commenting, editing, and writing to the manuscript. ES provided relevant resources. All authors contributed to the article and approved the submitted version.

Acknowledgments

We express our gratitude to Mathis Erichsen who developed parts of our data preparation R script. We would further like to thank Ethan R. Van Norman for sharing his R code on how

to calculate multilevel reliability and Chris Schatschneider for sharing detailed information on how he calculated multilevel reliability in his work. Finally, we acknowledge support from the Open Access Publication Fund of the University of Münster.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Bates, D., Martin, M., Ben, B., and Steve, W. (2015). Fitting linear mixed-effects models using LME4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Beisemann, M. (2022). A flexible approach to modelling over-, under- and equidispersed count data in IRT: The two-parameter conway-maxwell-poisson model. *Br. J. Math. Stat. Psychol.* 75, 411–443. doi: 10.1111/bmsp.12273
- Bollen, K. A. (1980). Issues in the comparative measurement of political democracy. *Am. Sociol. Rev.* 45:370. doi: 10.2307/2095172
- Brown, A., and Croudace, J. T. (2015). "Scoring and Estimating Score Precision Using Multidimensional IRT Models," in *Multivariate Applications Series. Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, eds P. Steven, Reise, A. Dennis, and Revicki (Routledge/Taylor & Francis Group), 307–333.
- Bulut, O., and Cormier, D. C. (2018). Validity evidence for progress monitoring with star reading: Slope estimates, administration frequency, and number of data points. *Front. Educ.* 3:68. doi: 10.3389/feduc.2018.00068
- Christ, T. J., and Desjardins, C. D. (2018). Curriculum-based measurement of reading: An evaluation of frequentist and bayesian methods to model progress monitoring data. *J. Psychoeduc. Assess.* 36, 55–73. doi: 10.1177/0734282917712174
- Christ, T. J., Johnson-Gros, K. N., and Hintze, J. M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychol. Sch.* 42, 615–622. doi: 10.1002/pits.20107
- Christ, T. J., Monaghan, B. D., Zopluoglu, C., and Van Norman, E. R. (2013a). Curriculum-based measurement of oral reading: Evaluation of growth estimates derived with pre-post assessment methods. *Assess. Effect. Interv.* 38, 139–153. doi: 10.1177/1534508412456417
- Christ, T. J., Zopluoglu, C., Long, J. D., and Monaghan, B. D. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Except. Children* 78, 356–373. doi: 10.1177/001440291207800306
- Christ, T. J., Zopluoglu, C., Monaghan, B. D., and Van Norman, E. R. (2013b). Curriculum-based measurement of oral reading: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *J. Sch. Psychol.* 51, 19–57. doi: 10.1016/j.jsp.2012.11.001
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Except. Children* 52, 219–232. doi: 10.1177/001440298505200303
- Deno, S. L. (1987). Curriculum-based measurement. *Teach. Except. Children* 20, 40–42. doi: 10.1177/004005998702000109
- DiStefano, C., and Zhu, M. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Pract. Assess. Res. Eval.* 14:20. doi: 10.7275/da8t-4g52
- Ferrando, P. J., and Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educ. Psychol. Measur.* 78, 762–780. doi: 10.1177/0013164417719308
- Förster, N., Erichsen, M., and Forthmann, B. (2021). Measuring reading progress in second grade: Psychometric properties of the Quop-L2 test series. *Eur. J. Psychol. Assess.* [Epub ahead of print]. doi: 10.1027/1015-5759/a000688
- Förster, N., and Kuhn, J.-T. (2021). Ice is hot and water is dry: Developing equivalent reading tests using rule-based item design. *Eur. J. Psychol. Assess.* [Epub ahead of print]. doi: 10.1027/1015-5759/a000691
- Forthmann, B., Förster, N., and Souvignier, E. (2022). Shaky student growth? a comparison of robust bayesian learning progress estimation methods. *J. Intell.* 10:16. doi: 10.3390/jintelligence10010016
- Forthmann, B., Gühne, D., and Doebl, P. (2020b). Revisiting dispersion in count data item response theory models: The conway-maxwell-poisson counts model. *Br. J. Math. Stat. Psychol.* 73, 32–50. doi: 10.1111/bmsp.12184
- Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., and Holling, H. (2020c). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *Br. J. Educ. Psychol.* 90, 683–699. doi: 10.1111/bjep.12325
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., and Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *J. Educ. Measur.* 21, 347–360. doi: 10.1111/j.1745-3984.1984.tb01039.x
- Haertel, E. H. (2006). "Reliability," in *Educational measurement*, ed. R. L. Brennan (Westport, CT: Praeger Publishers), 65110.
- Jorgensen, T. D., Sunthud, P., Alexander, M. S., and Yves, R. (2021). *SemTools: Useful Tools for Structural Equation Modeling*. Available online at: <https://cran.r-project.org/package=semTools> (accessed September 7, 2022).

- Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., and Fox, J. P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychol. Methods* 14, 54–75. doi: 10.1037/a0014877
- Maris, G., and van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika* 77, 615–633. doi: 10.1007/s11336-012-9288-y
- McMaster, K. L., Shin, J., Espin, C. A., Jung, P.-G., Wayman, M. M., and Deno, S. L. (2017). Monitoring elementary students' writing progress using curriculum-based measures: Grade and gender differences. *Read. Writ.* 30, 2069–2091. doi: 10.1007/s11145-017-9766-9
- National Center on Intensive Intervention (2014). *Progress monitoring technical review committee: Frequently asked questions*. American Institutes for Research. Available online at: https://intensiveintervention.org/sites/default/files/APM_FAQs_2014.pdf
- Parker, D. C., McMaster, K. L., Medhanie, A., and Silberglitt, B. (2011). "Modeling early writing growth with curriculum-based measures." *Sch. Psychol. Q.* 26, 290–304. doi: 10.1037/a0026833
- Pornprasertmanit, S., Miller, P., Schoemann, A., and Jorgensen, T. D. (2020). "Simsem: SIMulated Structural Equation Modeling." *R package version 0.5-15*. Available online at: <https://CRAN.R-project.org/package=simsem> (accessed September 7, 2022).
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear model: Applications and data analysis methods*, 2nd Edn. Los Angeles, CA: SAGE.
- Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *Br. J. Math. Stat. Psychol.* 54, 315–323. doi: 10.1348/000711001159582
- Rosseel, Y. (2012). Lavaan : An R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02
- Schatschneider, C., Wagner, R. K., and Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learn. Individ. Dif.* 18, 308–315. doi: 10.1016/j.lindif.2008.04.005
- Schurig, M., Jungjohann, J., and Gebhardt, M. (2021). Minimization of a short computer-based test in reading. *Front. Educ.* 6:684595. doi: 10.3389/feduc.2021.684595
- Silberglitt, B., and Hintze, J. M. (2007). How Much growth can we expect? a conditional analysis of r—cbm growth rates by level of performance. *Excep. Children* 74, 71–84. doi: 10.1177/001440290707400104
- Snijders, T. A. B., and Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd Edn. Los Angeles: Sage.
- Souvignier, E., Förster, N., Hebbeker, K., and Schütze, B. (2021). "Using Digital Data to Support Teaching Practice - Quop: An Effective Web-Based Approach to Monitor Student Learning Progress in Reading and Mathematics in Entire Classrooms," in *International Perspectives on School Settings, Education Policy and Digital Strategies. A Transatlantic Discourse in Education Research*, eds S. Jörnitz and A. Wilmers (Leverkusen: Budrich), 283–298.
- Thornblad, S. C., and Christ, T. J. (2014). Curriculum-based measurement of reading: Is 6 weeks of daily progress monitoring enough?" edited by christy walcott. *Sch. Psychol. Rev.* 43, 19–29. doi: 10.1080/02796015.2014.12087451
- Van Norman, E. R., Christ, T. J., and Zopluoglu, C. (2013). The effects of baseline estimation on the reliability, validity, and precision of cbm-r growth estimates. *Sch. Psychol. Quart.* 28, 239–255. doi: 10.1037/spq0000023
- Van Norman, E. R., and Parker, D. C. (2018). A comparison of split-half and multilevel methods to assess the reliability of progress monitoring outcomes. *J. Psychoeduc. Assess.* 36, 616–627. doi: 10.1177/0734282917696936
- Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002
- VanDerHeyden, A. M., and Burns, M. K. (2008). Examination of the utility of various measures of mathematics proficiency. *Assess. Effect. Interv.* 33, 215–224. doi: 10.1177/1534508407313482
- West, S. G., Taylor, A. B., and Wu, W. (2012). "Model Fit and Model Selection in Structural Equation Modeling," in *Handbook of Structural Equation Modeling*, ed. R. H. Hoyle (The Guilford Press), 209–231.
- Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educ. Measur.* 36, 52–61. doi: 10.1111/emip.12165
- Wise, S. L., and DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educ. Assess.* 15, 27–41. doi: 10.1080/10627191003673216



OPEN ACCESS

EDITED BY

Mohamed A. Ali,
Grand Canyon University,
United States

REVIEWED BY

Leanne R. Ketterlin Geller,
Southern Methodist University,
United States
Mariama Njie,
Grand Canyon University,
United States

*CORRESPONDENCE

Sven Anderson
sven.anderson@tu-dortmund.de

SPECIALTY SECTION

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

RECEIVED 15 May 2022

ACCEPTED 30 August 2022

PUBLISHED 28 November 2022

CITATION

Anderson S, Schurig M, Sommerhoff D and
Gebhardt M (2022) Students' learning
growth in mental addition and subtraction:
Results from a learning progress
monitoring approach.
Front. Psychol. 13:944702.
doi: 10.3389/fpsyg.2022.944702

COPYRIGHT

© 2022 Anderson, Schurig, Sommerhoff
and Gebhardt. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC
BY\)](#). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Students' learning growth in mental addition and subtraction: Results from a learning progress monitoring approach

Sven Anderson^{1*}, Michael Schurig¹, Daniel Sommerhoff² and Markus Gebhardt³

¹Faculty of Rehabilitation Sciences, TU Dortmund University, Dortmund, Germany, ²Department of Mathematics Education, IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany, ³Faculty of Human Sciences, University of Regensburg, Regensburg, Germany

The purpose of this study was to measure and describe students' learning development in mental computation of mixed addition and subtraction tasks up to 100. We used a learning progress monitoring (LPM) approach with multiple repeated measurements to examine the learning curves of second- and third-grade primary school students in mental computation over a period of 17 biweekly measurement intervals in the school year 2020/2021. Moreover, we investigated how homogeneous students' learning curves were and how sociodemographic variables (gender, grade level, the assignment of special educational needs) affected students' learning growth. Therefore, 348 German students from six schools and 20 classes (10.9% students with special educational needs) worked on systematically, but randomly mixed addition and subtraction tasks at regular intervals with an online LPM tool. We collected learning progress data for 12 measurement intervals during the survey period that was impacted by the COVID-19 pandemic. Technical results show that the employed LPM tool for mental computation met the criteria of LPM research stages 1 and 2. Focusing on the learning curves, results from latent growth curve modeling showed significant differences in the intercept and in the slope based on the background variables. The results illustrate that one-size-fits-all instruction is not appropriate, thus highlighting the value of LPM or other means that allow individualized, adaptive teaching. The study provides a first quantitative overview over the learning curves for mental computation in second and third grade. Furthermore, it offers a validated tool for the empirical analysis of learning curves regarding mental computation and strong reference data against which individual learning growth can be compared to identify students with unfavorable learning curves and provide targeted support as part of an adaptive, evidence-based teaching approach. Implications for further research and school practice are discussed.

KEYWORDS

learning progress monitoring, mathematics education, mental computation, latent growth curve model, continuous norming, learning progression, formative assessment, curriculum-based measurement (CBM)

Introduction

Mental computation can be defined as a person's ability to perform basic arithmetic operations correctly and quickly in their mind by using adequate solution strategies without resorting to external resources such as paper and pencil or a calculator (e.g., Maclellan, 2001; Varol and Farran, 2007). Focusing on current curricula, mental computation has an essential place in primary school mathematics education (e.g., the *Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany* (KMK), 2005; Seeley, 2005; *National Council of Teachers of Mathematics* (NCTM), 2022). This importance can be explained by the fact that mental computation has a high value in everyday life (e.g., Reys, 1984). Moreover, previous research has pointed to the great influence of mental computation for higher-order mathematical thinking (e.g., Blöte et al., 2000; Hickendorff et al., 2019; Pourdavood et al., 2020). In particular, mental computation can support students in understanding the concept of numbers, in discovering computational strategies, in making reasonable estimates and in developing a flexible and adaptive handling of these when solving mathematical problems. Furthermore, mental computation is a basis for written computation and its mastery.

Research findings indicate that most students improve their mental computation skills during primary school years and are able to solve multi-digit addition and subtraction tasks adequately in grades 3 or higher (e.g., Heirdsfield and Cooper, 2004; Karantzis, 2011). However, empirical research also shows that a large number of students struggle with mental computation throughout and beyond primary school (e.g., Reys et al., 1993; Miller et al., 2011; Hickendorff et al., 2019). Recent research findings (e.g., Peltenburg et al., 2012; Gebhardt et al., 2015; Rojo and Wakim, 2022) suggests, that the acquisition of multi-digit mental computation is particularly challenging for students with special educational needs (SEN). For example, students with SEN in the area of learning (SEN-L) mostly exhibit a lack of solid basic arithmetic skills, which is often responsible for difficulties and missing learning success in secondary school mathematics (e.g., Gebhardt et al., 2014; Rojo and Wakim, 2022). Studies focusing on the mathematical learning development of students with and without SEN conclude that students with SEN not only show a lower mathematical achievement, but also have a slower learning growth than their peers without SEN (e.g., Wei et al., 2013; Gebhardt et al., 2015).

In light of the importance of mental computation for further mathematical achievement and the high number of students who have difficulties developing adequate mental computation skills, there is great need for providing information about students' learning growth in educational research and practice (e.g., Salaschek et al., 2014). For teachers in particular, summative assessments at the beginning or end of the school year are often insufficient in identifying struggling students at an early stage. An alternative are formative assessments, which provide diagnostic information during the learning process and allows for

instructional adjustment (e.g., Cisterna and Gotwals, 2018). One formative approach is learning progress monitoring (LPM) which is discussed as an appropriate method to provide teachers with ongoing feedback on students' learning development (e.g., Deno et al., 2009). To evaluate learning growth with LPM tools, teachers regularly administer short parallel tests and assess students' individual learning curves using LPM graphs. The evaluation of these individual learning curves is the basis for decisions about maintaining or adjusting educational instructions. For example, within the Response to Intervention (RTI) approach, LPM tools are used to identify struggling students who would benefit from additive educational instruction or to evaluate the effectiveness of learning offers (e.g., Stecker et al., 2008).

In order to address the need for information on student learning development and learning growth in mental computation in educational research and practice, the purpose of the present study was to examine the latent learning curves as mean learning growth of the individual learning curves of second and third grade students in mental computation of mixed addition and subtraction tasks. Therefore, we used a recently developed computation test (Anderson et al., 2022). The present study first investigated the psychometric quality of this test for LPM. This included how the measures are related to student performance on standardized arithmetic tests and whether the LPM tool can sensitively measure student's learning and progress at different ability levels. Subsequently, we used the data to describe students' latent learning curves regarding mental computation skills in addition and subtraction over a period of 17 biweekly measurement intervals. Based on this, we examine differential developments using sociodemographic characteristics such as gender and grade level as well as the assignment of SEN.

Mental addition and subtraction and its differential development

Mental computation skills regarding basic arithmetic are an important prerequisite for the acquisition of mathematical literacy as measured in international school performance studies such as the Programme for International Student Assessment (PISA; Organisation for Economic Co-operation and Development (OECD), 2018). Moreover, these competencies are inherent to the primary school mathematics curricula. In particular, mastering mental computation of multi-digit addition and subtraction tasks is an important learning goal in primary school all over the world. According to primary school mathematics curricula of all federal states in Germany or in the United States (e.g., KMK, 2005; NCTM, 2022), students should have developed profound mental addition and subtraction skills in the number range up to 100 by the end of grade 2. Based on a spiral approach, mental addition and subtraction skills are extended to three-digit numbers at the beginning of grade 3. By the end of primary school, students should be able to transfer these skills to higher number ranges. Subsequently, in the second half of grade 3, students learn the

written algorithms for addition and subtraction of three-digit numbers (Selter, 2001). Considering the curricular requirements, third graders should therefore be able to routinely carry out two-digit mental addition and subtraction tasks whereas this may be more challenging for second graders.

The results of previous studies (e.g., Bryant et al., 2008; Karantzis, 2011) indicate that some students' performance in mental addition and subtraction is at a low level even at higher grades, implying urgent need for educational means to address this issue. Weak performance in this area in primary school is attributed to different task characteristics that contribute to task difficulty (e.g., Benz, 2003, 2005) and the use of inefficient solving strategies (e.g., Beishuizen, 1993; Cooper et al., 1996; Beishuizen et al., 1997; Heirdsfield and Cooper, 2004; Varol and Farran, 2007).

Regarding the task characteristics, the construction of the numbers, whose sum value or difference value needs to be calculated, plays an important role. Research has shown that multi-digit addition and subtraction tasks vary in their difficulty and probability of solving them correctly (e.g., Benz, 2003, 2005). This is explained by the fact that there are multiple difficulty-generating item characteristics (DGICs) that have an influence on task difficulty (e.g., the number of digits of a term or the necessity of crossing ten). Knowledge about the influence of different DGICs is particularly important for rule-based item design of school achievement tests (e.g., for statistical word problems see Holling et al., 2009). For mathematical word problems, Daroczy et al. (2015) provide a review of DGICs that contribute to the difficulty of such tasks. Anderson et al. (2022) discuss the advantages of rule-based item design and the identification of DGICs for constructing a pool of items for a mixed addition and subtraction test for LPM.

Besides that, the flexible and adequate use of different solution strategies for solving multi-digit addition and subtraction tasks is relevant (for an overview, e.g., Torbeyns et al., 2009; Hickendorff et al., 2019). While solving single-digit addition and subtraction tasks is based on the retrieval of the solution from long-term memory as an arithmetic fact, the outcome of multi-digit addition and subtraction tasks must be computed based on the adaptive application of known solution strategies. With the use of inefficient solution strategies such as counting strategies, multi-digit addition and subtraction tasks are solved slowly and often incorrectly. In addition to the flexibility in choosing appropriate solution strategies in correspondence with the requirement of a specific task, hurdles for struggling students include a lack of the conceptual understanding of numbers and a lack of fluency in using computation procedures (e.g., Verschaffel et al., 2007).

With regard to students' solving strategies of multi-digit addition and subtraction tasks, two complementary dimensions can be distinguished respecting number-based strategies: the operation that is necessary for the solution process and the way the numbers are used in the solution process (Hickendorff et al., 2019). Concerning the first dimension, multi-digit addition only allows direct addition, while multi-digit subtraction allows several options (direct subtraction, indirect addition, indirect

subtraction). Concerning the second dimension, there are different strategies to manipulate numbers to successfully master the computation process. In sequencing strategies, numbers are interpreted as objects on a mental number line and addition is seen as moving forward and subtraction as moving backward on it. For example, the addition task $44 + 38$ is given. The direct addition with the sequencing strategy would be computed as $44 + 30 = 74$; $74 + 8 = 82$. In decomposition strategies, numbers are interpreted as objects with a decimal structure and the operations require splitting or portioning the numbers. With the decomposition strategy it would be computed as $40 + 30 = 70$; $4 + 8 = 12$; $70 + 12 = 82$. In varying situations, different strategies are used that adaptively consider both the numbers and the operations in the solution process. These two complementary dimensions can be used to categorize students' problem-solving strategies. Students with mathematical difficulties often have problems acquiring the different strategies and using them in an adaptive and flexible way. These students use inefficient solution strategies (e.g., counting strategies) and are often unable to accurately solve single-digit addition and subtraction, which is a prerequisite for successful acquisition of multi-digit strategy skills (e.g., Varol and Farran, 2007; Verschaffel et al., 2007).

Despite the high curricular importance, previous qualitative studies already indicate a large heterogeneity in the development of two-digit addition and subtraction computation skills during the second school year (e.g., Benz, 2003, 2007). Previous research has also shown that students with SEN have difficulty acquiring adequate computation skills (Gersten et al., 2005; Evans, 2007; Bryant et al., 2008; Wei et al., 2013; Soares et al., 2018). For example, at the end of primary school, many students with SEN-L have acquired lower competencies in the development in mathematics in general (e.g., Gebhardt et al., 2014, 2015) and in the development of mental arithmetic computation for numbers up to 100 compared to their peers without SEN-L (Peltenburg et al., 2012; Rojo and Wakim, 2022).

While research findings on the difficulties for students with SEN-L are consistent, this is not the case for gender-based performance differences in mental computation (e.g., Winkelmann et al., 2008; Wei et al., 2013; Pina et al., 2021). Wei et al. (2013) reported significant and persistent gender performance differences in favor of boys among students with different SEN that persisted from primary to secondary school. For regular primary education, Pina et al. (2021) found no significant gender differences in mathematics achievement in computation. Results from international large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS) indicated gender-based differences in average mathematics achievement between girls and boys at the end of primary school. In TIMSS 2019, fourth grade boys showed a higher average performance than girls in almost half of the 58 participating countries. In four countries, girls had a higher average achievement than boys. In 27 countries, gender equity of average performance in mathematics was reported. For the arithmetic domain, boys achieved higher test scores than girls in

almost all countries and for more than half of the countries' differences are even significant (Mullis et al., 2020). In Germany, the differences in this domain are significant (Nonte et al., 2020).

In contrast, in a study of third-through eighth-graders with and without SEN, Yarbrough et al. (2017) found statistically significant differences between boys and girls in favor of girls in grades 5, 7, and 8 for learning growth in mental computation. The tests included mathematical computation tasks on the four basic arithmetic operations. The difficulty of the tasks varied according to the respective curricular requirements of the respective grade. However, the knowledge about students' differential latent learning curves when acquiring mental computation skills is limited. For example, the results on learning growth by Yarbrough et al. (2017) were based on only three measurement time points. As noted above, there is only a small number of longitudinal surveys, including a large number of measures for the valid assessment of latent learning curves.

Learning progress monitoring

Due to heterogeneous student learning, there is an increasing need for teachers to use data about individual student's learning development for their instructional decision-making (e.g., Espin et al., 2017). In this regard, LPM is a promising method that provides data on individual students' learning development and assists teachers identifying learning problems in early stages as well as in evaluating the achievement of learning goals. One approach of LPM is curriculum-based measurement (CBM; e.g., Deno, 1985; Stecker et al., 2005): a set of procedures that can be used frequently and quickly to assess student learning progress and the effectiveness of instruction in academic domains such as reading, spelling, writing, or computation (e.g., Hosp et al., 2016). CBM procedures consist of short parallel tests that require only a few minutes (e.g., 1–5 min) and items are typically based on the identification of robust indicators or on curriculum sampling (Fuchs, 2004). Finding robust indicators includes identifying tasks that best represent the various subskills of a specific domain or that correlate strongly with them. For reading, oral reading fluency is regarded as a robust indicator of general reading competence and comprehension (Deno et al., 1982). In the domain of mathematics, number sense is considered a robust indicator for mathematics performance in kindergarten and first grade primary school (e.g., Lembke and Foegen, 2009). Curriculum sampling involves selecting exemplary tasks that assess curricular learning goals. Each CBM test is then aligned with curricular objectives that are relevant to the entire assessment period (e.g., Fuchs, 2004). With regard to highly heterogeneous learning groups, for example in inclusive classrooms, strictly curriculum-based LPM are of limited use because students with SEN (e.g., SEN-L) are often not taught according to the regular class curriculum (Gebhardt et al., 2016).

Results of LPM usually output a sum score (e.g., number of correctly solved tasks) and the learning development is

represented in a graph. To represent individual learning development, linear trends at the student level are often estimated. Therefore, the parameters intercept and slope are relevant. The slope represents the mean learning growth of a student (e.g., the proportion of additional tasks that were solved correctly in the comparison of the measurement points). The intercept contains information about the approximated individual learning level at the beginning of LPM. For reliable and valid conclusions about learning development, Christ et al. (2013) recommend using data from at least six measurements. Based on these data, teachers can then decide whether the instruction used promotes learning success as intended (individual learning curve is as expected), whether the instruction used should be adjusted (individual learning curve is lower than expected), or whether the learning goal can be adjusted because the individual learning curve is higher than initially expected (e.g., Espin et al., 2017).

Since the 1970s, a large body of LPM research has focused on the development and application of instruments for different domains with a focus on reading (for an overview see Tindal, 2013). Until 2008, LPM research mostly focused on the domain of reading and not mathematics (e.g., Van Der Heyden and Burns, 2005; Foegen et al., 2007). In a review of the literature concentrating on the development of LPM in mathematics, Foegen et al. (2007) found that only a small part of the studies focused on mathematics and here primarily on preschool and elementary mathematics. In German-speaking countries, LPM research has advanced in recent years, especially in educational psychology and special education, addressing several academic domains, including reading and math (for an overview see Breitenbach, 2020; Gebhardt et al., 2021).

Learning progress monitoring of mathematics computation

Regarding different types of LPM in mathematics, Hosp et al. (2016) differentiate between tests for number sense (early numeracy), for computational skills (computation), and for the application of mathematical skills such as interpreting measurements, tables, or graphs (concepts and applications). For LPM in the area of computational skills, there are differences in how the tasks are intended to be solved (e.g., in a mental or written way) and how large the assessment domain is in each case (assessment of a single skill or multiple skills). According to Christ et al. (2008), the domain of mathematics computation is especially suitable for a frequently used LPM tool that can be used in research as well as data-based instructional decision-making. Instruments for this domain are usually constructed to provide very brief measurements of a relatively narrow arithmetic performance range and LPM tasks corresponding to the curricular level or individual learning objectives. There is also evidence that teachers can use the data of computation LPM to improve the performance of students with SEN (for an overview see Foegen et al., 2007).

LPM of (mental) computational skills does not aim to measure mathematical literacy as in PISA (OECD, 2018), and it does not address the language requirements of number word problems, which can also play a role in understanding mathematics. In contrast, it focuses on (mental) computation skills as an important prerequisite for solving word problems as well as mathematical literacy in general (e.g., Varol and Farran, 2007). Still, this narrow focus must be considered when selecting potential criterion measures for the evaluation of criterion validity. According to Christ et al. (2008), the coefficients of criterion validity between LPM tools and standardized mathematical achievement tests that measure overall performance in mathematics can therefore only be interpreted to a limited extent, whereas criterion validity is understandably much higher for procedures that relate exclusively to arithmetic tasks or include subtests in the domain of computation.

A variety of LPM tools for computation and mental computation have been developed in the past decades, especially in the United States (for an overview, e.g., Christ et al., 2008; Tindal, 2013). In German-speaking countries, some tools have been established for LPM (mental) computation. For example, Sikora and Voß (2017) have developed and empirically validated a curriculum-based LPM tool for the four basic arithmetic skills for grades 3 and 4. In composing the LPM tests, they considered item characteristics that may influence item difficulty (e.g., number range, arithmetic operation, digits to be computed in item solution, place value, and standard form tasks). The LPM tests of Strathmann and Klauer (2012) or Salaschek and Souvignier (2014) have integrated mental computation tasks as part of a broader curriculum-based LPM tool. These are usually a subset of a few items, each testing one of the four basic arithmetic competencies at the respective curricular level. Anderson et al. (2022) developed a test based on an item-generating system for mixed addition and subtraction tasks for numbers up to 100. This test is built on multiple difficulty-generating item characteristics (DGICs). First, three DGICs were deduced from prior mathematics education research (arithmetic operation, necessity of crossing ten, the number of second term digits) and varied within the item design process so that all possible combinations were adequately represented in an item pool. Subsequently the Rasch model (RM) and the Linear Logistic Test Model (LLTM) were used to estimate and predict the influence of the DGICs. The results of the LLTM approach indicate that all three suspected difficulty-generating characteristics were significant predictors of item difficulty and explain about 20% of the variance in the item difficulty parameters of the RM. Results suggest that DGICs can influence item difficulty across grade levels and ensure long-term use across multiple grade levels. Thus, identified curriculum-independent DGICs have the potential to be used to construct LPM tests for classes with curriculum-independent learners. In test development, the present study follows the item generation system reported by Anderson et al. (2022) and extends it to include an additional DGIC.

Requirements for learning progress monitoring

In order to validly assess learning progress, frequent and regular use of LPM requires a large number of parallel tests that should be mostly consistent in difficulty and are sufficiently sensitive to measure learning. Therefore, LPM tools have to address a variety of psychometric properties. This includes classical test quality criteria (e.g., validity, reliability) as well as psychometric criteria such as one-dimensionality, homogeneous test difficulty, sensitivity to change, and test fairness (e.g., Wilbert and Linnemann, 2011; Schurig et al., 2021). For example, identifying characteristics that have an influence on task difficulty can support the development of parallel tests with homogeneous test difficulty (e.g., Wilbert, 2014; Anderson et al., 2022). In this regard, LPM tests should be constructed under the assumptions of item response theory (IRT), which features sample independence, non-linear dependencies between trait and response, and the ability to test multiple parameters of response behavior (e.g., Schurig et al., 2021). For the practical purpose of data-based decision-making, the results should also be as easy as possible for teachers to interpret and use to choose or adapt instruction (e.g., Espin et al., 2017). In particular, computer- or web-based LPM tools can contribute to improving the usability in schools through a high degree of automation of test generation and evaluation (e.g., Mühling et al., 2019).

As evidence for its use in progress measurement, Fuchs (2004) proposed a three-stage systematization of LPM research. Research at stage 1 includes studies that aim to test the psychometric adequacy of the tool as a status diagnostic. Stage 2 includes all research that provides evidence that a LPM tool can sensitively and validly represent learning growth over time. Research at stage 3 involves studies that examine whether the use of LPM data for instructional decisions improves student performance. For all academic domains, a large part of the prior research has focused on stage 1 and addressed the psychometric adequacy of LPM tool as a status diagnostic (Fuchs, 2017).

Purpose of the study

The purpose of the present study was to examine the latent learning curves of second and third grade primary school students in mental addition and subtraction with a newly developed web-based LPM tool. Our study thus addresses stages 1 and 2 outlined by Fuchs (2004). Therefore, we examined the psychometrical adequacy of the LPM tool at an individual measurement point as well as its sensitivity to learning growth over time by addressing the following research questions:

Research question 1.1: How do the LPM test scores at different measurement time points relate to standardized school achievement test results at the beginning and the end of the survey period?

Research question 1.2: How reliable are the results of our LPM tool in terms of correlations between different measurement time points?

Research question 1.3: Is the LPM tool sensitive to student learning at different ability levels?

Building on these analyses, we subsequently examined students' latent learning curves regarding mental addition and subtraction in second and third grade over a period of 17 biweekly measurement intervals, focusing on the overall learning development over time as well as interindividual heterogeneity therein. As prior results have highlighted that sociodemographic characteristics can influence learning and learning development, we additionally examined, if gender and grade level as central sociodemographic characteristics as well as the assignment of SEN lead to empirically distinguishable learning curves. Research questions are as follows:

Research question 2.1: How homogeneous are students' latent learning curves over a period of 17 biweekly measurement intervals?

Research question 2.2: Do students' latent learning curves differ between groups with different sociodemographic characteristics such as gender and grade level?

Research question 2.3: What influence does the assignment of SEN have on students' latent learning curves in mental computation?

Materials and methods

Participants and setting

A total of 348 students from nine second-grade and nine third-grade inclusive education classes and two third-grade special education classes¹ of six schools participated in the study. The schools were located in urban as well as rural areas of North Rhine-Westphalia as state of the Federal Republic of Germany. The six schools were recruited by convenience sampling. Therefore, it was taken into account that a similar number of second and third graders participated in the survey, as well as students in a special school and students in inclusive schools. Of the participating students, 162 (46.55%) were in the second grade, 186 (53.45%) in the third grade. The average age of students at the start of the study was 8.43 years ($SD=0.80$). Further sociodemographic characteristics of the participating students at the start of the study are reported in Table 1. A number of 38 students (10.92%) had the assignment of SEN, most of them in the area of learning (SEN-L:

TABLE 1 Sociodemographic of students at the start of the study.

Personal characteristics	Full sample		Grade 2		Grade 3	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Gender						
Female	176	50.57	85	52.47	91	48.92
Male	172	49.43	77	47.53	95	51.08
SEN						
Yes	38	10.92	2	1.23	36	19.35
No	310	89.08	160	98.77	150	80.65
Migration background						
Yes	96	27.59	46	71.60	50	26.88
No	252	72.41	116	28.40	136	73.12

19; 50.00% of the SEN students) or in the area of communication and interaction (SEN-CI: 17; 44.74% of the SEN students).

Measures and procedure

The study was conducted from November 2020 to July 2021 and covered a period of 17 biweekly measurement intervals. At the beginning and at the end of the survey period, arithmetic subscales of the standardized German paper-pencil test DEMAT 2+ (German Mathematics Test for Second Grade and for the beginning of Third Grade; Krajewski et al., 2020) were administered. The DEMAT 2+ is representative of all German regular second-grade mathematics curricula and is suitable as a norm-based test for the last months of the second and the first months of the third school year. The test contains tasks for numbers up to 100. For this study, we selected subscales of the DEMAT 2+ that included computation tasks without mathematics word problems. These included tasks for number properties, addition and subtraction place values tasks, tasks for doubling and halving numbers, and tasks for calculating with money (see Table 2). The use of DEMAT 2+ subscales at the beginning were followed by LPM every 2 weeks. At the end of the survey period, DEMAT 2+ subscales were administered a second time. Credit was given only for completely correct answers.

The used LPM tool included mixed addition and subtraction tasks for numbers up to 100, which required students to enter the correct solution (for addition tasks the sum value; for subtraction tasks the difference value) into a blank field. We designed the items using a rule-based approach that considered several DGICs derived in advance from mathematics education research and evaluated in Anderson et al. (2022). Extending the results of Anderson et al. (2022) four DGICs were used to model the difficulty of the items: the arithmetic operation (addition versus subtraction; DGIC 1), the necessity of crossing ten (no crossing versus with crossing; DGIC 2), the number of second term digits (one-digit numbers versus two-digit numbers; DGIC 3), and the necessity to add up to the next full ten (not necessary versus necessary; DGIC 4). Based on these four DGICs, we created a pool of 3,027 items. The four DGICs were varied within the item design

¹ In inclusive education classes in Germany, students with and without SEN are taught together. Students with SEN sometimes have individual learning goals that do not have to correspond to the curricular goals of classmates without SEN. In special education classes, only students with SEN are taught. The learning goals can follow curricular or individual learning objectives, depending on the type of SEN.

TABLE 2 Subscales of the DEMAT 2+ (Krajewski et al., 2020) used for this study.

Subscale DEMAT 2+	Requirement	Example	No. of items
Number properties	Identification of even and odd two-digit numbers	Identify the even numbers! 25 44 8 19 8 38 17	2
Addition place value	Identification of the correct first/second summand	Calculate! ... + 15 = 34	4
Subtraction place value	Identification of the correct subtrahend/minuend	Calculate! 56 - ... = 36	4
Doubling numbers	Doubling of a two-digit number (with and without crossing ten)	Take the double! 70 → ...	3
Halving numbers	Halving of a two-digit number (with and without crossing ten)	Take the half! 24 → ...	3
Calculating with money	Calculation of a two-digit cent amount to get 1 € (1 € = 100 cents)	How many cents are missing if you want 1€? At 45 cents missing ...	4

process so that all possible combinations were adequately represented in the item pool (see Table 3).

The item pool was implemented on an online platform² (Gebhardt et al., 2016; Mühling et al., 2019). Based on an equal distribution of the 10 possible item categories in the item selection, a fixed order was established for the baseline test. Starting at the second measurement, items were drawn from the total item pool in a randomized, however equally distributed manner according to the 10 item categories. Students could not skip any drawn items during the test time. We assume that an equal distribution of the items on the described item categories causes a harmonization of the difficulties of the tests. Based on this, an individual test was created for each student by the online platform for each additional measurement. Accordingly, from the second measurement on, we assume missing completely at random (MCAR) for all non-drawn items.

Trained administrators tested students in their classrooms in groups of 5–10 during class time. To perform the test, each participating student used a tablet device. Testing time was 5 min. The students had to mentally compute the tasks without external support. At the beginning of each measurement, students received a short technical briefing, sample tasks were solved, and students had the opportunity to ask the test administrator questions. Students could then start the test themselves by clicking on a start button. Tests ended automatically after 5 min testing time. In the time allotted, students were instructed to answer as many mathematics computation tasks as possible. Each probe contains a substantial number of tasks, making it unlikely that a student could finish within the time limit. No partial credit was given for partially correct answers.

All students who participated in at least one LPM test were included in the following analyses. Not all students participated in LPM at each measurement. The main reasons for this were home schooling periods during the survey due to the COVID-19 pandemic, a staggered start to the surveys within the participating schools, individual absence of students, or technical problems. In order to compute the latent mean-growth and comparable, time-dependent norms across the survey time, 17 equidistant measurement intervals were derived from the raw data. In order

TABLE 3 Sample items illustrating different types of items based on the four DGICs.

Category	Example	DGIC 1	DGIC 2	DGIC 3	DGIC 4
1	27 + 2	Addition	No	One	No
2	23 + 13	Addition	No	Two	No
3	21 + 9	Addition	No	One	Yes
4	52 + 38	Addition	No	Two	Yes
5	78 + 9	Addition	Yes	One	No
6	67 + 27	Addition	Yes	Two	No
7	48 - 3	Subtraction	No	One	No
8	98 - 24	Subtraction	No	Two	No
9	65 - 7	Subtraction	Yes	One	No
10	91 - 16	Subtraction	Yes	Two	No

DGIC 1: Arithmetic operation; DGIC 2: Crossing ten; DGIC 3: No. of 2nd term digits; DGIC 4: Add up to the next full ten.

to establish reasonable distance interval lengths to observe change, 2 weeks were chosen as the length of the interval. Due to practical reasons within the schools, some children were tested twice within one interval and not within another. When students were tested twice, only the first observation within a measurement interval was used. Data are available for 12 of the 17 biweekly measurement intervals, the other intervals are missing due to homeschooling and holidays.

Statistical analyses

Participation

The individual number of participation related to the LPM measurement intervals in this study varied ($M = 5.98$; $SD = 2.20$). The range of participation is 10, with 24 students (6.90%) of the total sample participating in the surveys only once and 5 students (1.44%) participating 11 times. 282 students (81.03%) participated in at least five, 222 students (63.79%) in at least six LPM measurement intervals.

Analyses

The presentation of descriptive statistics is followed by the results on the research questions. To address research question 1.1, criterion and predictive validity were analyzed by examining how

² www.levumi.de

LPM scores relate to the employed arithmetic subscale scores of the standardized paper-pencil mathematics test DEMAT 2+. To answer research question 1.2, a RM was fitted for every single LPM test. Due to the high number of missing data by design, the item fit was evaluated using a conditional pairwise item category comparison implemented in the *R* package *pairwise* (Heine and Tarnai, 2015). The pairwise approach is able to handle (completely) random missing data by design. Subsequently, alternate form test-retest reliability for adjacent and more distant tests was calculated. Regarding research question 1.3, performance is assessed by the continuous norming method using the *R* package *cNorm* (Lenhard et al., 2018) to evaluate how sensitive the test measures at different ability levels at different measurement intervals. In *cNorm*, norm values and percentiles are estimated as a function of time and possibly covariates using Taylor polynomials. To identify adequate test norms, a polynomial regression model needs to be found that describes the norming sample as accurately as possible with the minimum number of predictors. Lenhard and Lenhard (2021) emphasized that higher numbers of terms do often lead to overfit. Therefore, *cNorm* used $k=4$ terms by default. In the modeling process the stopping criterion is $R^2=0.99$.

In our case, the explanatory variable represents the different measurement intervals over the LPM survey period of 17 biweekly measurement intervals. Thus, the *cNorm* approach addresses some disadvantages of traditional norming methods such as a high sample size, the consideration of sampling errors or any distributional assumptions. Moreover, gaps between discrete levels of the explanatory variable can be closed (Gary et al., 2021). This can be particularly advantageous for LPM, since norm tables can be generated not only for the discrete measurement point of the

survey, but also for each subsequent measurement point, even if no measurement has occurred. In our case, this means that norm values could also be derived for the measurement intervals where no LPM tests were conducted due to homeschooling.

In order to address research questions 2.1–2.3, latent learning curves and the modeling of individual differences in learning growth over time including sociodemographic characteristics such as gender or grade and assignment of SEN are examined via latent growth curve modeling (LGCM; e.g., Muthen and Khoo, 1998). In educational and psychological contexts, this approach is often used to determine learning growth and the influence of background variables in LPM longitudinal data (e.g., Salaschek and Souvignier, 2014; Johnson et al., 2020). The lavaan package (Rosseel, 2012) was used to estimate latent growth curve models.

The LGCM illustrates the use of slope and intercept as two latent variables to model differences over time. The student's initial performance in solving mixed addition and subtraction tasks for numbers up to 100 is represented as a scale score (intercept). Similarly, the rate of linear growth in the student's competences across all measurement intervals is represented as a scale score (slope). The initial LGCM (Model 1) represented in Figure 1, includes each biweekly administration of LPM, except for measurement intervals 4–7 and 11 when no measurements could be taken in schools due to the COVID-19 pandemic and the switch to home schooling. Furthermore, it was analyzed if sociodemographic variables such as gender or grade level or the assignment of SEN influence learning growth. For this, the LGCM was extended to include group differences (Model 2). We used gender (0 = male, 1 = female), grade level (0 = grade 2, 1 = grade 3), and special educational need (0 = no, 1 = yes) as dummy coded variables across

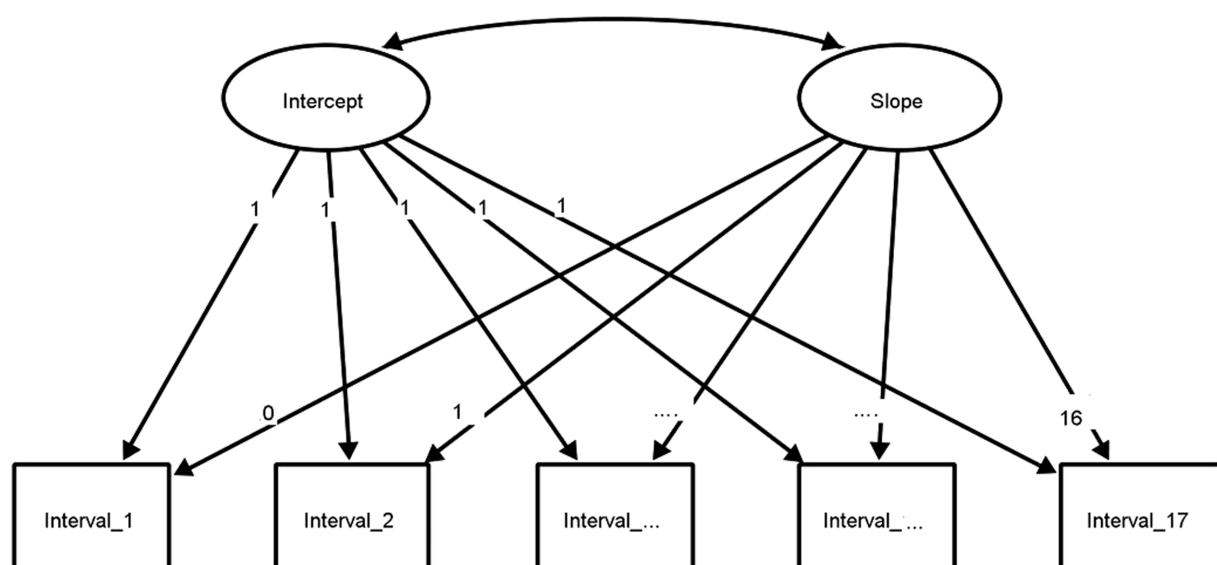


FIGURE 1
Graphical representation of a LGCM for 17 measurement intervals without covariates. The three dots represent the intervals 3–15, which were not included in this representation for greater clarity. The graphical representation of the growth model was created with Ω nyx (von Oertzen et al., 2015).

the 12 measurement intervals. In Model 2, the intercept and slope variables are predicted while considering these background variables.

Results

Descriptive statistics for the LPM tests at each of the 12 measurement intervals are presented in Table 4 for the full sample and separately for both grade levels. With regard to the measurement intervals 1–3 and 14–17, regular mathematics instruction took place at school. In contrast, the measurement intervals in between were often characterized by home schooling due to the COVID-19 pandemic when mathematical instruction often took place *via* distance learning and not all students were able to regularly participate in LPM testing.

Research question 1.1: Criterion validity

For reliability analysis, Cronbach's alpha and Mc Donald's omega were calculated to assess the internal consistency of the

subscales of the selected DEMAT 2+ at the first and last measurement time point separately for grades 2 and 3 and the full sample. The internal consistency of the subscales of the DEMAT 2+ are satisfactory (see Tables 5, 6). Correlations of LPM sum scores with the overall sum scores of the subscales of the DEMAT 2+ at the first measurement time point were strong with a mean correlation of $r=0.73$ (95%CI [0.68; 0.78]). For the full sample, the correlations of the various subscales of the DEMAT 2+ ranged from 0.39 (subscale number properties) to 0.67 (subscale calculating with money) with $M=0.57$ and $SD=0.10$.

At the last measurement time point, correlations of LPM sum scores with the overall sum scores of the DEMAT 2+ subscales were moderate with a mean correlation of $r=0.57$ (95%CI [0.49; 0.64]). For the full sample, the correlations of the various DEMAT 2+ subscales ranged from 0.22 (subscale number properties) to 0.57 (subscale addition place value) with $M=0.42$ and $SD=0.12$ (for further information separately by grade level see Table 7).

To test the predictive validity of LPM measures, the correlation of the LPM sum scores at the first measurement time point with the overall sum scores of the DEMAT 2+ subscales at the last measurement time point were calculated. Correlations were

TABLE 4 Descriptive statistics of LPM scores for each measurement interval.

Time of measurement	Full sample		Grade 2		Grade 3	
	<i>n</i>	<i>M</i> (<i>SD</i>)	<i>n</i>	<i>M</i> (<i>SD</i>)	<i>n</i>	<i>M</i> (<i>SD</i>)
Measurement interval 1	194	11.01 (7.80)	85	5.62 (4.00)	109	15.20 (7.46)
Measurement interval 2	256	10.24 (8.09)	131	6.00 (5.28)	125	14.69 (8.15)
Measurement interval 3	108	13.46 (9.46)	43	7.79 (6.59)	65	17.22 (9.23)
[...]	[...]	[...]	[...]	[...]	[...]	[...]
Measurement interval 8	150	12.21 (9.30)	89	9.47 (8.01)	61	16.20 (9.66)
Measurement interval 9	225	13.30 (10.02)	109	9.03 (7.91)	116	17.32 (10.17)
Measurement interval 10	152	15.11 (10.38)	81	11.27 (8.20)	71	19.49 (10.92)
[...]	[...]	[...]	[...]	[...]	[...]	[...]
Measurement interval 12	43	16.16 (11.27)	31	13.52 (10.33)	12	23.00 (11.10)
Measurement interval 13	46	16.91 (11.96)	34	13.56 (10.16)	12	26.42 (11.91)
Measurement interval 14	232	15.44 (9.46)	88	11.19 (7.90)	144	18.04 (9.41)
Measurement interval 15	291	15.65 (9.99)	133	11.97 (8.42)	158	18.75 (10.18)
Measurement interval 16	275	15.65 (10.26)	108	10.68 (6.65)	167	18.87 (10.90)
Measurement interval 17	107	17.42 (10.26)	59	14.59 (7.65)	48	20.90 (11.95)

LPM tests were canceled due to the COVID-19 pandemic in measurement intervals 4–7 and 11. Square brackets with three dots represent the canceled measurement intervals.

TABLE 5 Cronbach's Alpha and Mc Donald's Omega coefficients at first measurement time point.

DEMAT 2+ subscale	Full sample		Grade 2		Grade 3	
	α	ω	α	ω	α	ω
Number properties (2 items)	0.78	0.78	0.78	0.78	0.77	0.77
Addition place value (4 items)	0.80	0.80	0.71	0.73	0.76	0.77
Subtraction place value (4 items)	0.73	0.74	0.70	0.71	0.69	0.70
Doubling numbers (3 items)	0.88	0.89	0.81	0.83	0.90	0.91
Halving numbers (3 items)	0.72	0.73	0.62	0.65	0.72	0.73
Calculating w. money (4 items)	0.90	0.90	0.86	0.87	0.89	0.89

α = Cronbach's Alpha; ω = Mc Donald's Omega; Full sample = 328 students (grade 2 = 155; grade 3 = 173).

TABLE 6 Cronbach's Alpha and Mc Donald's Omega coefficients at last measurement time point.

DEMAT 2+ subscale	Full sample		Grade 2		Grade 3	
	α	ω	α	ω	α	ω
Number properties (2 items)	0.76	0.76	0.66	0.66	0.82	0.82
Addition place value (4 items)	0.80	0.80	0.73	0.74	0.78	0.78
Subtraction place value (4 items)	0.61	0.62	0.57	0.58	0.58	0.63
Doubling numbers (3 items)	0.84	0.86	0.82	0.85	0.86	0.88
Halving numbers (3 items)	0.78	0.80	0.74	0.75	0.80	0.83
Calculating w. money (4 items)	0.89	0.89	0.86	0.86	0.89	0.89

α = Cronbach's Alpha; ω = Mc Donald's Omega; Full sample = 302 students (grade 2 = 130; grade 3 = 172).

TABLE 7 Correlations of LPM scores at the beginning and end of the survey with subscales of the DEMAT 2+.

Variables of DEMAT 2+	Full sample		Grade 2		Grade 3	
	LPM Begin.	LPM End	LPM Begin.	LPM End	LPM Begin.	LPM End
Beginning of survey						
Number properties	0.39**	0.23**	0.27**	0.10	0.41**	0.22**
Addition place value	0.63**	0.47**	0.49**	0.28**	0.53**	0.43**
Subtraction place value	0.57**	0.52**	0.38**	0.37**	0.56**	0.51**
Doubling numbers	0.54**	0.42**	0.37**	0.40**	0.54**	0.34**
Halving numbers	0.63**	0.45**	0.44**	0.34**	0.62**	0.41**
Calculation w. money	0.67**	0.53**	0.46**	0.40**	0.63**	0.48**
Overall sum score subscales	0.73**	0.56**	0.55**	0.44**	0.70**	0.53**
End of survey						
Number properties	0.36**	0.22**	0.22*	0.10	0.41**	0.24**
Addition place value	0.58**	0.57**	0.22*	0.38**	0.60**	0.60**
Subtraction place value ^a	0.47**	0.45**	0.28**	0.39**	0.47**	0.43**
Doubling numbers	0.36**	0.33**	0.28**	0.33**	0.39**	0.31**
Halving numbers	0.52**	0.47**	0.30**	0.32**	0.61**	0.51**
Calculation w. money	0.52**	0.47**	0.35**	0.44**	0.52**	0.42**
Overall sum score subscales	0.63**	0.57**	0.40**	0.49**	0.66**	0.55**

* indicates $p < 0.5$. ** indicates $p < 0.01$. ^aThe subscale did not reach an acceptable internal consistency (see Table 6).

moderate to strong with a mean correlation of $r = 0.63$ (95%CI [0.56; 0.70]); for grade 2: $r = 0.40$, for grade 3: $r = 0.66$). Correlations of DEMAT 2+ sum scores at the first and at the last measurement time point were strong with a mean correlation of $r = 0.80$ (95%CI [0.75; 0.84]; for grade 2: $r = 0.61$, for grade 3: $r = 0.86$).

Research question 1.2: Reliability

The reliability of the resulting Weighted Maximum Likelihood Estimation (WLE) person parameters ranged from 0.80 to 0.85 ($M = 0.82$; $SD = 0.02$) for the measurement intervals. Furthermore, the alternate form test–retest reliability was calculated for each pair of adjacent and more distant tests (e.g., LPM interval 1 scores to LPM interval 2 scores, LPM interval 1 scores to LPM interval 17 scores, ..., LPM interval scores 16 to LPM interval scores 17; see Table 8). Correlation indices between scores from adjacent measurement intervals ranged from 0.73 to 0.93. With reference to the COTAN review system for evaluating test quality (Evers

et al., 2015), we interpret this as sufficient alternate form test–retest reliability.

Research question 1.3: Generating continuous tests norms

As mentioned above, the procedure is robust to different or small sample sizes. The modeling procedure of the LPM scores from interval 1 to interval 17 reached an adjusted $R^2 = 0.98$ with 5 terms and an intercept. It must be taken into account that at five measurement intervals no data collection could be conducted due to homeschooling and therefore $R^2 = 0.99$ was not reached. To achieve this value, the number of terms would have to be increased further, which we have refrained to avoid an overfit. The norms in the upper range vary strongly. Figure 2 shows the assignment of the raw test values at the various levels to a specific percentile. Students with high raw scores at the beginning also have a higher slope over the survey period. The clustering of percentiles in the

TABLE 8 Correlations of LPM sum scores.

	MI 1	MI 2	MI 3	[...]	MI 8	MI 9	MI 10	[...]	MI 12	MI 13	MI 14	MI 15	MI 16
MI 1													
MI 2	0.80*** [0.73, 0.86]												
MI 3	0.83*** [0.71, 0.90]	0.86*** [0.79, 0.91]											
[...]													
MI 8	0.83*** [0.67, 0.91]	0.85*** [0.79, 0.89]	0.89*** [0.82, 0.93]										
MI 9	0.77*** [0.67, 0.84]	0.78*** [0.72, 0.83]	0.79*** [0.68, 0.86]		0.84*** [0.79, 0.89]								
MI 10	0.89*** [0.78, 0.94]	0.85*** [0.79, 0.89]	0.86*** [0.77, 0.91]		0.87*** [0.82, 0.91]	0.83*** [0.76, 0.87]							
[...]													
MI 12		0.87*** [0.77, 0.93]	0.89*** [0.77, 0.95]		0.93*** [87, 0.96]	0.95*** [0.90, 0.97]	0.90*** [0.82, 0.95]						
MI 13		0.86*** [0.74, 0.92]	0.88*** [0.75, 0.94]		0.86*** [0.75, 0.92]	0.90*** [0.82, 0.94]	0.88*** [0.79, 93]		0.87*** [0.77, 0.93]				
MI 14	0.76*** [0.68, 0.82]	0.80*** [0.73, 0.85]	0.79*** [0.70, 0.86]		0.84*** [0.76, 0.89]	0.86*** [0.81, 0.89]	0.90*** [0.84, 0.93]		0.86*** [0.76, 0.93]	0.87*** [0.77, 0.93]			
MI 15	0.66*** [0.56, 0.74]	0.74*** [0.68, 0.79]	0.72*** [0.60, 0.80]		0.82*** [0.76, 87]	0.78*** [0.72, 0.83]	0.86*** [0.81, 0.90]		0.91*** [0.83, 0.95]	0.87*** [0.78, 0.93]	0.86*** [0.82, 0.89]		
MI 16	0.70*** [0.61, 0.77]	0.70*** [0.63, 0.76]	0.76*** [0.65, 0.84]		0.74*** [0.64, 0.81]	0.71*** [0.63, 0.78]	0.76*** [0.67, 0.83]		0.94*** [0.75, 0.99]	0.88*** [0.60, 0.97]	0.77*** [0.70, 0.82]	0.81*** [0.76, 0.85]	
MI 17	0.64** [0.29, 0.84]	0.79*** [0.71, 0.86]	0.74*** [0.49, 0.87]		0.74*** [0.64, 0.82]	0.70*** [0.59, 79]	0.85*** [0.79, 0.90]		0.91*** [0.67, 0.98]	0.85*** [0.54, 0.96]	0.81*** [0.67, 0.90]	0.85*** [0.78, 0.89]	0.81*** [0.73, 0.87]

MI is the acronym for the term measurement interval. LPM tests were canceled due to the COVID-19 pandemic in measurement intervals 4–7 and 11. Values in square brackets indicate the 95% confidence interval for each correlation. Square brackets with three dots represent the canceled measurement intervals. *indicates $p < 0.05$; **indicates $p < 0.01$; ***indicates $p < 0.001$. The correlation of MI 1 to M 12 and MI 1 to MI 13 cannot be reliably calculated due to a low sample size.

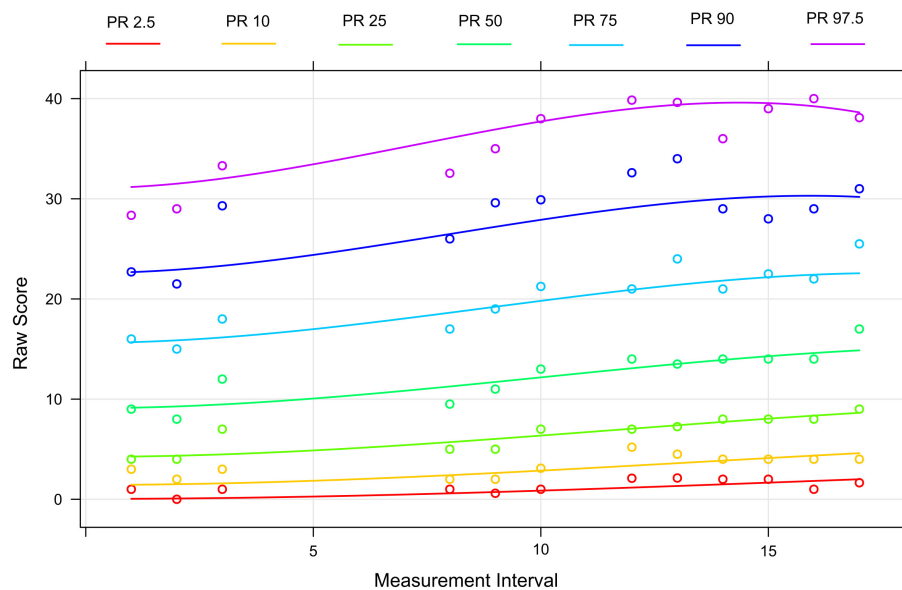


FIGURE 2

Percentile curves based on the sample of the mixed addition and subtraction LPM test. The curves show, which raw score (y-axis) is assigned to a specific ability level (each represented by a percentile curve) at a certain LPM measurement interval (x-axis).

lower ranges (roughly up to the 25th percentile) do indicate a low separability. In other words: the test is probably still too difficult for as many as 25% of the students.

Research questions 2.1–2.3: Sensitivity to learning

The investigation of students' latent learning curves in mental addition and subtraction is presented in two steps. In a first step, we report the model fit of Model 1 (model without covariates) and Model 2. In a second step, we evaluate the latent learning curves regarding each of the research questions 2.1–2.3.

To estimate the model fit, we used the chi-square test, the root mean square error of approximation (RMSEA), the Tucker-Lewis Index (TLI), the Comparative fit index (CFI), and the standardized root mean square residual (SRMR). TLI and CFI values close to 0.95 indicate an adequate fit to the data. RMSEA values close to 0.06 and SRMR values close to 0.08 generally are recommended (Hu and Bentler, 1999).

A first LGCM was estimated (Model 1) to investigate the changes in the means of the test scores over the measurement intervals. Model estimation terminated successfully for Model 1, $\chi^2(73) = 195.116$. The RMSEA for model 1 is 0.069, 90%CI [0.058, 0.081] which implies an adequate fit. The TLI for Model 1 is 0.953 and above the value for determining a good fit for model acceptability. The Comparative fit index (CFI) for Model 1 is 0.949. The standardized root mean square residual (SRMR) is 0.068.

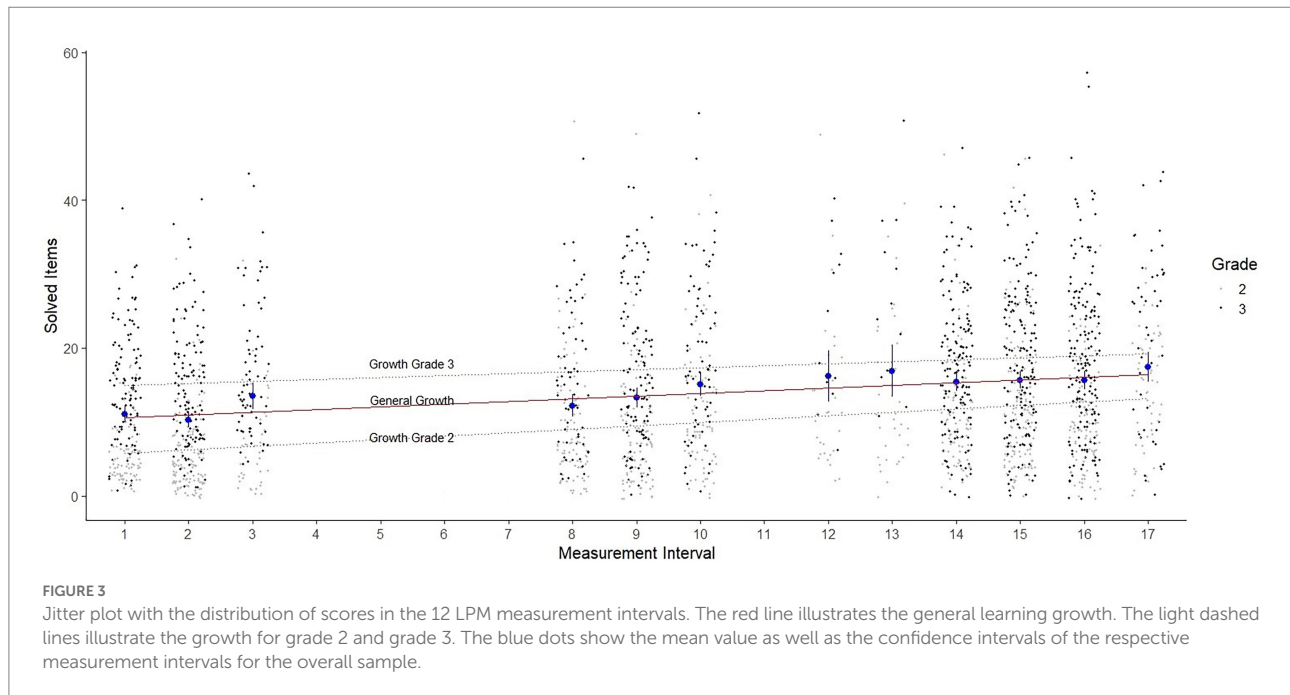
The slope, as a measure of linear growth in mental addition and subtraction competence over time, is positive for Model 1

(estimate = 0.342; $SE = 0.024$; $p < 0.001$), indicating that mental computation skills have improved over the survey period (see also Figure 3). On average, the students solved roughly one more task correctly every three measurement intervals. Considering the grade level, second grade students solved roughly one more task correctly every two measurement intervals (estimate = 0.439; $SE = 0.030$; $p < 0.001$), for third grade students this was about every four measurement intervals (estimate = 0.249; $SE = 0.033$; $p < 0.001$). Data thus suggest a slightly steeper learning curve for second graders, implying faster learning.

The variance of the slope is also statistically significant for Model 1 ($p < 0.001$), indicating that learning growth did not change at the same rate for all students (see also Figure 3). Of the 222 students who completed the minimum six measurement time points required by Christ et al. (2013), 213 students (95.95%) had an individual positive slope, indicating that they exhibited learning growth over time. Positive slope values ranged from 0.001 to 1.176, indicating that some students were able to solve up to one more task per interval on average.

In the previous LGCM model (Model 1), individual change over time was indicated by intercept and slope, including only grade as covariate. In a further step, we extend the LGCM model to include group differences according to the research questions 2.2 and 2.3.

Model estimation terminated successfully for Model 2, $\chi^2(113) = 236.477$. The RMSEA for model 2 is 0.056, 90% CI [0.046, 0.066] which implies a close to adequate fit. The TLI for model 2 is 0.952 and is also above the value for determining a good fit for the model acceptability. For Model 2, CFI is 0.952 and the SRMR is 0.057. In comparison to Model 1, the indices suggest a slightly better fit of Model 2.



The intercept of the latent learning curves in model 2 differed based on gender (estimate gender = -3.313 , $p < 0.001$). Data revealed a higher intercept for males in comparison to females at the beginning of the measurement, that is male participants were able to solve approximately three tasks more correctly than females. The intercept also differed based on grade level (estimate grade level = 10.311 , $p < 0.001$), indicating that third graders solved ~ 10 tasks more than second graders at the beginning of the measurement. Furthermore, the intercept differed based on the assignment of SEN (estimate SEN-L = -9.385 , $p < 0.001$; estimate SEN-CI = -4.015 , $p = 0.009$). This indicates that students with special educational needs in the area of learning solved ~ 9 tasks less than students without such special need, whereas students with a special need in the area of communication and interaction solved ~ 4 tasks less. All results are reported in Table 9.

Focusing on the impact of the factors on the learning slope, only grade level led to a significantly differing learning slope (estimate = -0.178 , $p < 0.001$), indicating that third graders learning was slightly slower than second graders learning. The learning slope did not significantly differ for males versus females (estimate = -0.073 , $p = 0.105$) or for students with and without SEN-L (estimate = -0.150 , $p = 0.139$) or with and without SEN-CI (estimate = -0.017 , $p = 0.868$).

Discussion

The present study used a newly developed LPM tool to investigate the latent learning growth curves in mental addition and subtraction of second and third graders and the influence of sociodemographic characteristics such as grade level, gender, and the assignment of SEN on these curves. Thus, this study addressed

TABLE 9 Parameter estimates for linear latent growth model (Model 2).

		Estimate	Estimate (Std. all)	<i>p</i>
<i>Mean</i>				
Intercept		3.644	0.478	0.187
	Slope	1.077	3.503	≤ 0.001
<i>Variance</i>				
Intercept		29.758	0.512	≤ 0.001
	Slope	0.083	0.879	≤ 0.001
<i>Covariances</i>				
Intercept–slope		0.531	0.338	≤ 0.001
<i>Regressions</i>				
Intercept	SEN-L	-9.385	-0.280	≤ 0.001
	SEN-CI	-4.015	-0.114	0.009
	Grade level	10.311	0.675	≤ 0.001
	Gender	-3.313	-0.217	≤ 0.001
Slope	SEN-L	-0.150	-0.111	0.139
	SEN-CI	-0.017	-0.012	0.868
	Grade level	-0.178	-0.289	≤ 0.001
	Gender	-0.073	-0.119	0.105

the research stages 1 and 2 outlined by Fuchs (2004) to classify LPM research, both of which are prerequisites for the valid interpretation of the gathered data regarding individual learning curves.

LPM research at stage 1

In order to address research stage 1, we used a number of reliability and validity tests to examine the psychometric quality

of the LPM tool as a static score. Correlations between sum scores of adjacent measurement intervals were strong, while sum scores of measurement intervals more distant in time showed the expected, somewhat lower correlations. As measured criterion validity, correlations between the LPM sum scores at the first and last measurement point with the arithmetic subscales of the DEMAT 2+ were moderate to strong. As expected, the correlations were lower at the last measurement time point. In this regard, it should be kept in consideration that the DEMAT 2+ reflects requirements of the mathematics curriculum of the second grade (Krajewski et al., 2020). By the end of grade 3, most students without SEN should be able to solve the items of the DEMAT 2+. Furthermore, as a measure of predictive validity, the association between the LPM sum scores at the first measurement and sum scores of the DEMAT 2+ arithmetic subscales were moderate to strong and are an indication of the important role of a mental computation in the solution of further arithmetic problems. Thus, even a single measurement in winter with the LPM tool can be a solid predictor of arithmetic performance at the end of the school year.

LPM research at stage 2

Over and above the psychometric characteristics at stage 1, significant positive linear growth in LGCM analyses indicates that the LPM tool is sensitive to students learning (Stage 2). Both, the slopes and the variance in slopes were significant, showing that meaningful learning has occurred over the 17 measurement intervals and that students significantly differ in their learning growth. This is also reflected in the broad range of individual slope values. These findings are consistent with the results of the study by Salaschek and Souvignier (2014). In their study, they reported significant differences in learning growth in second grade students' computation skills. In their study, second graders on average solved just under one more item per 3-week measurement interval, whereas in our study students solved one more item correctly every 4 weeks. Nevertheless, the results are only comparable to a limited extent as the LPM computation tests by Salaschek and Souvignier (2014) included tasks with all four basic arithmetic operations and reflected second-grade curriculum goals. In contrast, the LPM test employed in this study included mixed addition and subtraction tasks with varying difficulty based on the underlying DGICs. Moreover, the LPM test in this study required students to write the correct solution in a blank field, which allows a qualitative analysis of errors and eliminates guessing, the LPM computation tests by Salaschek and Souvignier (2014) were presented in a multiple-choice format.

Regarding the comparison of learning growth for weaker and stronger students, based on the continuous norming approach, we observed that students in the upper percentiles have higher learning growth than students in the lower percentiles, who barely improved over the measurement intervals. This highlights

prior longitudinal or crossed-lagged findings regarding the high impact of prior knowledge on future learning in mathematics (e.g., Star et al., 2009) and underlines the relevance of this research. However, analyses also highlight a positive result: Almost 96% of the students achieved an individual positive slope even though the positive slope values were relatively heterogeneous ranging from 0.001 to 1.176 (i.e., an average improvement between 0.001 and 1.176 items over the whole measurement period of 17 biweekly measurement intervals). Nonetheless, the results for the growth curves show a significant floor effect for students at the lower end of the distribution. These findings are of particular practical relevance, as it highlights the benefit of close use of LPM tools to identify learners with small or no learning growth at an early stage and provide appropriate learning support to prevent learning stagnation and ongoing mathematical difficulties. Therefore, heterogeneity in classes should be increasingly reflected in instructional decisions (e.g., Stecker et al., 2008).

In addition to these results, our study also provides information on the influence of sociodemographic characteristics such as gender and grade or the assignment of SEN on learning growth in mental computation. In our study, we found significant differences in participants' prior achievement in favor of students in higher grades and students without SEN. Moreover, there are also differences in students learning growth development. In particular, the higher learning growth for second graders is consistent with curricular expectations and results of previous research (e.g., Selter, 2001; Benz, 2005; Karantzis, 2011). In the second grade, mental addition and subtraction with one- and two-digit numbers is curricularly established and taught, whereas in the third grade, there is already an emphasis on the written computational algorithm and some students already have a fairly high level of mental computational skills. We found gender differences for the intercept, but not for the slope. Students with SEN had a significantly lower intercept value. This result is consistent with the findings that especially students with SEN-L often do not master the basic arithmetic operations taught in primary school even in secondary school (e.g., Peltenburg et al., 2012; Gebhardt et al., 2014; Rojo and Wakim, 2022).

Limitations

There are some limitations to consider in our study. First, the COVID-19 pandemic played an important role even before the survey began (e.g., home schooling as early as the 2019/2020 school year), which implies that the results should not be interpreted free of these home schooling influences. The COVID-19 pandemic also resulted in the cancellation of scheduled measurements due to homeschooling during this survey. As a result, it was not possible to carry out all the planned measurements at all six participating schools. This resulted in a smaller than expected amount of data being available for some measurement intervals. Moreover, the observed latent learning

curves may be somewhat less steep than expected with regular teaching. Thus, future longitudinal surveys will need to confirm our findings.

Second, only few students with the assignment of SEN participated in the study and they were unevenly distributed across the grades. This is mainly due to the fact that in Germany, SEN, especially SEN-L, is often not allocated until the third grade.

Third, while mental computation is an important domain of overall mathematics competence, it is also a relatively narrow focus in regard of mathematics skills. Therefore, it is not appropriate to interpret the results in such a way that they provide valid information about the overall performance in mathematics (e.g., see Christ et al., 2008).

Fourth, the influence of other important individual characteristics on mathematics performance such as working memory or language skills were not addressed in our study, although these could have an influence on task processing (e.g., Purpura and Ganley, 2014).

Fifth, our mental computation test consisted of visually administered items. Previous research (e.g., Reys et al., 1995) suggests that students' mental computation performance may be influenced by the mode of task presentation (e.g., visually or orally). This cannot be investigated in our study as the test did not contain orally administered items.

Sixth, the results show that the tasks are suitable for measuring learning development, but do not yet cover all performance domains. In particular, more simple computation tasks are needed to more accurately measure learning development in the lower skill range in the future. For this purpose, the used DGICs can provide valuable information about the obstacles to solving tasks correctly and for the construction of easier tests that can more sensitively measure mental computation skills at the lower performance levels.

Seventh, our study does not provide information about solution strategies that students used when completing the multi-digit addition and subtraction tasks. Accordingly, no statements can be made about the adequacy and flexibility of the students' use of solution strategies. Nevertheless, we assume that a higher sum value of correct items over time implies a more elaborate use of solution strategies.

Future research

Future studies need to further investigate how LPM tests can be systematically used by teachers to improve the mental computation skills of their students. Identifying where differences in mental computation occur can support teachers develop appropriate educational instruction to meet the needs of individual students (e.g., Yarbrough et al., 2017). Our item design based on four DGICs will allow us to make statements that are even more concrete about areas that were specifically challenging for students, possibly pointing to student misconceptions and thus

area that need specific teacher attention and support. In this regard, we will be able to offer not only a general performance score, but also differentiated scores according to the four DGICs. This allows us to provide teachers with more specific qualitative feedback on students' mental computational performance. In the context of DGIC-focused analyses, there are several questions of relevance: A first important question would be whether the influence of DGICs changes over time (e.g., whether the DGIC necessity of crossing ten loses influence over time). For example, in order to provide tailored math instruction, for teachers it would be useful to know which hurdles in the learning process students have already successfully mastered and which they have not. Following on from this, a second important question is which students have longer-term difficulties in mastering specific hurdles. Our results show that in particular students with SEN have lower skills and less learning growth over time. A further investigation could be to examine the reasons for these performance differences and apparent stagnation of some low-achieving students, which for example might be related to insufficient knowledge or ineffective use of specific computation strategies.

Furthermore, future studies should examine trajectories in mental computation to describe how students differ in their skills and what characterizes different groups of learners. This information can both help identify students with learning difficulties in mental computation and provide trajectory-specific instructions (e.g., Salaschek et al., 2014).

Another issue arises from the construction of the parallelized tests that we used. While they were parallel in item selection based on the DGICs and should thus be comparable regarding their difficulty, there is no specific way to test this hypotheses. However, we assume, that the randomization by item category harmonize the difficulties enough to observe substantial inference.

In conclusion, we developed an LPM tool for mental computation that meets the criteria of LPM research stages 1 and 2. This lays important foundations for its future use as an LPM instrument in general as well as in regard of its use in computerized adaptive testing approaches (e.g., Frey and Seitz, 2009). However, to normalize scores that address a broader proficiency range by computerized adaptive testing, the scoring mechanism (e.g., sum scores) has to be modified and the item parameters have to be fixed. We believe that the current study is a step in this direction.

The results of our study underline the high variability of mental computation skills and illustrate that one-size-fits-all instruction is not appropriate. Instead, teachers need to obtain insight into the different learning growth curves based on LPM data and provide individualized learning offers (e.g., Hickendorff et al., 2019). Otherwise, a lack of mental computation skills can be a hurdle for future learning success in mathematics. The study provides a strong reference against which individual growth can be compared to identify struggling students in mental computation and provide targeted support based on qualitative error analysis.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

SA is the primary author, conducted data collection, data preparation and analysis, created the initial version of the manuscript, and guided the further writing process. MS supported data collection, data preparation and analysis, and provided feedback in the writing process. DS provided theoretical expertise and feedback in the writing process. MG provided writing oversight and feedback in the writing process. All authors contributed to the article and approved the submitted version.

Funding

The current research is part of the project Dortmund Profile for Inclusion-Oriented Learning and Teacher Training

References

- Anderson, S., Sommerhoff, D., Schurig, M., Ufer, S., and Gebhardt, M. (2022). Developing learning progress monitoring tests using difficulty-generating item characteristics: an example for basic arithmetic operations in primary schools. *J. Educ. Res. Online* 14, 122–146. doi: 10.31244/jero.2022.01.06
- Beishuizen, M. (1993). Mental strategies and materials or models for addition and subtraction up to 100 in Dutch second grades. *J. Res. Math. Educ.* 24, 294–323. doi: 10.2307/749464
- Beishuizen, M., van Putten, C. M., and van Mulken, F. (1997). Mental arithmetic and strategy use with indirect number problems up to one hundred. *Learn. Instr.* 7, 87–106. doi: 10.1016/S0959-4752(96)00012-6
- Benz, C. (2003). "Irgendwie habe ich mir das aus dem Kopf geholt: Vorgehensweise von Zweitklässlern bei Additions- und Subtraktionsaufgaben im Hunderterraum am Schuljahresbeginn["Somehow I got it out of my head." Second graders' approach to addition and subtraction tasks for numbers up to hundred at the beginning of the school year"]" in *Beiträge zum Mathematikunterricht 2003*. ed. H. W. Henn (Franzbecker: Hildesheim), 101–104.
- Benz, C. (2005). *Erfolgsquoten, Rechenmethoden, Lösungswege und Fehler von Schülerinnen und Schülern bei Aufgaben zur Addition und Subtraktion im Zahlenraum bis 100* [Students' success rates, calculation methods, solutions and mistakes in addition and subtraction tasks in the range up to 100]. Hildesheim: Franzbecker.
- Benz, C. (2007). Die Entwicklung der Rechenstrategien bei Aufgaben des Typs $ZE \pm ZE$ im Verlauf des zweiten Schuljahres [The development of computational strategies in $ZE \pm ZE$ type tasks in the second year of primary school]. *J. Math. Didakt.* 28, 49–73. doi: 10.1007/BF03339333
- Blöte, A. W., Klein, A. S., and Beishuizen, M. (2000). Mental computation and conceptual understanding. *Learn. Instr.* 10, 221–247. doi: 10.1016/S0959-4752(99)00028-6
- Breitenbach, E. (2020). *Diagnostik. Eine Einführung* [Diagnostics. An introduction]. Wiesbaden: Springer.
- Bryant, D. P., Bryant, B. R., Gersten, R., Scamacca, N., and Chavez, M. M. (2008). Mathematics intervention for first- and second-grade students with mathematics difficulties. *Remedial Spec. Educ.* 29, 20–32. doi: 10.1177/0741932507309712
- Christ, T. J., Scullin, S., Tolbize, A., and Jiban, C. L. (2008). Implications of recent research: curriculum-based measurement of math computation. *Assess. Eff. Interv.* 33, 198–205. doi: 10.1177/1534508407313480
- Christ, T. J., Zopluoglu, C., Monaghan, B. D., and van Norman, E. R. (2013). Curriculum-based measurement of oral reading: multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *J. Sch. Psychol.* 51, 19–57. doi: 10.1016/j.jsp.2012.11.001
- Cisterna, D., and Gotwals, A. W. (2018). Enactment of ongoing formative assessment: challenges and opportunities for professional development and practice. *J. Sci. Teach. Educ.* 29, 200–222. doi: 10.1080/1046560X.2018.1432227
- Cooper, T. J., Heirdsfield, A., and Irons, C. J. (1996). "Children's mental strategies for addition and subtraction word problems," in *Children's Number Learning*. eds. J. T. Mulligan and M. C. Mitchelmore (Adelaide: Australian Association of Mathematics Teachers and Mathematics Education Research Group of Australasia), 147–162.
- Daroczy, G., Wolska, M., Meurers, W. D., and Nuerk, H.-C. (2015). Word problems: a review of linguistic and numerical factors contributing to their difficulty. *Front. Psychol.* 6:348. doi: 10.3389/fpsyg.2015.00348
- Deno, S. L. (1985). Curriculum-based measurement: the emerging alternative. *Except. Child.* 52, 219–232. doi: 10.1177/001440298505200303
- Deno, S. L., Mirkin, P. K., and Chiang, B. (1982). Identifying valid measures of reading. *Except. Child.* 49, 36–45. doi: 10.1177/001440298204900105

– DoProfil. DoProfil is part of the 'Qualitätsoffensive Lehrerbildung', a joint initiative of the Federal Government and the Länder, which aims to improve the quality of teacher training. The programme is funded by the Federal Ministry of Education and Research (Bundesministerium für Forschung und Bildung; Förderkennzeichen 01JA1930). The authors are responsible for the content of this publication.

Acknowledgments

We acknowledge financial support by Deutsche Forschungsgemeinschaft and Technische Universität Dortmund/TU Dortmund University within the funding programme Open Access Costs.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., et al. (2009). Developing a school-wide progress-monitoring system. *Psychol. Schs.* 46, 44–55. doi: 10.1002/pits.20353
- Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., and de Rooij, M. (2017). Data-based decision-making: developing a method for capturing teachers' understanding of CBM graphs. *Learn. Disabil. Res. Pract.* 32, 8–21. doi: 10.1111/ldrp.12123
- Evans, D. (2007). Developing mathematical proficiency in the Australian context: implications for students with learning difficulties. *J. Learn. Disabil.* 40, 420–426. doi: 10.1177/00222194070400050501
- Evers, A., Lucassen, W., Meijer, R., and Sijtsma, K. (2015). COTAN review system for evaluating test quality. Available at: <https://www.psynip.nl/wp-content/uploads/2019/05/NIP-Brochure-Cotan-2018-correctie-1.pdf> (Accessed August 11, 2022).
- Foegen, A., Jiban, C. L., and Deno, S. L. (2007). Progress monitoring measures in mathematics. *J. Spec. Educ.* 41, 121–139. doi: 10.1177/00224669070410020101
- Frey, A., and Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement: current state and future challenges. *Stud. Educ. Eval.* 35, 89–94. doi: 10.1016/j.stueduc.2009.10.007
- Fuchs, L. S. (2004). The past, present and future of curriculum-based measurement research. *Sch. Psychol. Rev.* 33, 188–192. doi: 10.1080/02796015.2004.12086241
- Fuchs, L. S. (2017). Curriculum-based measurement as the emerging alternative: three decades later. *Learn. Disabil. Res. Pract.* 32, 5–7. doi: 10.1111/ldrp.12127
- Gary, S., Lenhard, W., and Lenhard, A. (2021). Modelling norm scores with the cNORM package in R. *Psych* 3, 501–521. doi: 10.3390/psych3030033
- Gebhardt, M., Diehl, K., and Mühlhling, A. (2016). Online Lernverlaufsmessung für alle SchülerInnen in inklusiven Klassen [Online learning progress monitoring for all students in inclusive classes. www.LEVUMI.de]. *Zeitschrift für Heilpädagogik* 67, 444–453.
- Gebhardt, M., Jungjohann, J., and Schurig, M. (2021). *Lernverlaufsdiagnostik im förderorientierten Unterricht: Testkonstruktionen, Instrumente, Praxis* [Learning progress monitoring in remedial education: test construction, instruments, practice]. München: Ernst Reinhardt.
- Gebhardt, M., Sälzer, C., Mang, J., Müller, K., and Prenzel, M. (2015). Performance of students with special educational needs in Germany: findings from programme for international student assessment 2012. *J. Cogn. Educ. Psych.* 14, 343–356. doi: 10.1891/1945-8959.14.3.343
- Gebhardt, M., Zehner, F., and Hessels, M. (2014). Basic arithmetical skills of students with learning disabilities in the secondary special schools: an exploratory study covering fifth to ninth grade. *FLR* 2, 50–63. doi: 10.14786/flr.v2i1.73
- Gersten, R., Jordan, N., and Flojo, J. R. (2005). Early identification and intervention for students with mathematics difficulties. *J. Learn. Disabil.* 38, 293–304. doi: 10.1177/00222194050380040301
- Heine, J.-H., and Tarnai, C. (2015). Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychol. Test Assess. Model.* 57, 3–36.
- Heirdsfield, A. M., and Cooper, T. J. (2004). Factors affecting the process of proficient mental addition and subtraction: case studies of flexible and inflexible computers. *J. Math. Behav.* 23, 443–463. doi: 10.1016/j.jmathb.2004.09.005
- Hickendorff, M., Torbeyns, J., and Verschaffel, L. (2019). “Multi-digit addition, subtraction, multiplication, and division strategies,” in *International Handbook of Mathematical Learning Difficulties*. eds. A. Fritz, V. G. Haase and P. Räsänen (Cham: Springer International Publishing), 543–560.
- Holling, H., Bertling, J. P., and Zeuch, N. (2009). Automatic item generation of probability word problems. *Stud. Educ. Eval.* 35, 71–76. doi: 10.1016/j.stueduc.2009.10.004
- Hosp, M. K., Hosp, J. L., and Howell, K. W. (2016). *The ABC's of CBM: A Practical Guide to Curriculum-Based Measurement*. 2nd ed. New York: The Guilford Press.
- Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model. Multidiscip. J.* 6, 1–55. doi: 10.1080/1070519909540118
- Johnson, K. N., Thompson, K. L., and Farmer, R. L. (2020). Determining growth sensitivity of Star math with a latent growth curve model. *Can. J. Sch. Psychol.* 35, 197–209. doi: 10.1177/08295735202922678
- Karantzis, I. (2011). Mental arithmetic calculation in the addition and subtraction of two-digit numbers. The case of third and fourth grade elementary school pupils. *Int. J. Math. Educ.* 3, 3–24. Available at: <https://eclass.upatras.gr/modules/document/file.php/PDE1308/3%CE%BF%20%CE%86%CF%81%CE%B8%CF%81%CE%BF.pdf> (Accessed August 11, 2022).
- KMK (2005). *Bildungsstandards im Fach Mathematik für den Primarbereich: Beschluss der Kultusministerkonferenz der Länder der Bundesrepublik Deutschland vom 15.10.2004* [Educational standards in mathematics for primary education: resolution of the conference of the ministers of education and cultural affairs of the Länder in the federal republic of Germany from 15.10.2004]. Neuwied: Luchterhand.
- Krajewski, K., Dix, S., and Schneider, W. (2020). *DEMAT 2+; Deutscher Mathematiktest für zweite Klassen [DEMAT 2+; German mathematics test for second grade and for the beginning of third grade]*. 2nd ed. Göttingen: Hogrefe.
- Lembke, E. S., and Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learn. Disabil. Res. Pract.* 24, 12–20. doi: 10.1111/j.1540-5826.2008.01273.x
- Lenhard, W., and Lenhard, A. (2021). Improvement of norm score quality via regression-based continuous norming. *Educ. Psychol. Meas.* 81, 229–261. doi: 10.1177/0013164420928457
- Lenhard, W., Lenhard, A., and Gary, S. (2018). cNorm: Continuous norming [R package]. Available at: <https://cran.r-project.org/web/packages/cNORM/> (Accessed April 26, 2022).
- MacLellan, E. (2001). Mental calculation: its place in the development of numeracy. *Westminst. Stud. Educ.* 24, 145–154. doi: 10.1080/0140672010240205
- Miller, S. P., Stringfellow, J. L., Kaffar, B. J., Ferreira, D., and Mancl, D. B. (2011). Developing computation competence among students who struggle with mathematics. *Teach. Except. Child.* 44, 38–46. doi: 10.1177/004005991104400204
- Mühlhling, A., Jungjohann, J., and Gebhardt, M. (2019). “Progress monitoring in primary education using Levumi: a case study,” in *CSEDU 2019. Proceedings of the 11th International Conference on Computer Supported Education, 2–4 May, 2019, Heraklion, Greece*. Eds. H. Lane, S. Zvacek, and J. Uhomobih (SCITEPRESS-Science and Technology Publications), 137–144.
- Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., and Fishbein, B. (2020). *TIMSS 2019: International Results in Mathematics and Science*. Chestnut Hill: International Study Center, Lynch School of Education, Boston College.
- Muthén, B. O., and Khoo, S.-T. (1998). Longitudinal studies of achievement growth using latent variable modeling. *Learn. Individ. Differ.* 10, 73–101. doi: 10.1016/S1041-6080(99)80135-6
- NCTM (2022). Principles and standards: Number and operations, national council of teachers of mathematics. Available at: <https://www.nctm.org/Standards-and-Positions/Principles-and-Standards/Number-and-Operations/> (Accessed April 26, 2022).
- Nonte, S., Steinmayr, R., and Scholz, L. A. (2020). “Geschlechterunterschiede in mathematischen und naturwissenschaftlichen Kompetenzen [Gender differences in mathematics and science competencies]” in *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich [Mathematical and Scientific Competences of Primary School Students in Germany in International Comparison]*. eds. K. Schwippert, D. Kasper, O. Köller, N. McElvany, C. Selzer and M. Steffensky et al. (Waxmann: Münster), 223.
- OECD (2018). *PISA for Development Assessment and Analytical Framework*. Paris: OECD.
- Peltenburg, M., van den Heuvel-Panhuizen, M., and Robitzsch, A. (2012). Special education students' use of indirect addition in solving subtraction problems up to 100 – a proof of the didactical potential of an ignored procedure. *Educ. Stud. Math.* 79, 351–369. doi: 10.1007/s10649-011-9351-0
- Pina, V., Martella, D., Chacón-Moscote, S., Saracostti, M., and Fenollar-Cortés, J. (2021). Gender-based performance in mathematical facts and calculations in two elementary school samples from Chile and Spain: an exploratory study. *Front. Psychol.* 12:703580. doi: 10.3389/fpsyg.2021.703580
- Pourdavoud, R., McCarthy, K., and McCafferty, T. (2020). The impact of mental computation on children's mathematical communication, problem solving, reasoning, and algebraic thinking. *Athens J. Educ.* 7, 241–254. doi: 10.30958/aje.7-3-1
- Purpura, D. J., and Ganley, C. M. (2014). Working memory and language: skill-specific or domain-general relations to mathematics? *J. Exp. Child Psychol.* 122, 104–121. doi: 10.1016/j.jecp.2013.12.009
- Reys, R. E. (1984). Mental computation and estimation: past, present, and future. *Elem. Sch. J.* 84, 547–557. doi: 10.1086/461383
- Reys, B. J., Reys, R. E., and Hope, J. A. (1993). Mental computation: a snapshot of second, fifth and seventh grade student performance. *Sch. Sci. Math.* 93, 306–315. doi: 10.1111/j.1949-8594.1993.tb12251.x
- Reys, R. E., Reys, B. J., Nohda, N., and Emori, H. (1995). Mental computation performance and strategy use of Japanese students in grades 2, 4, 6, and 8. *J. Res. Math. Educ.* 26, 304–326. doi: 10.2307/749477
- Rojo, M., and Wakim, N. (2022). Teaching whole number addition and subtraction to students with learning disabilities. *Interv. Sch. Clin.* 10534512221081240. doi: 10.1177/10534512221081240
- Rosseel, Y. (2012). Lavan: an R package for structural equation modeling. *J. Stat. Soft.* 48, 1–36. doi: 10.18637/jss.v048.i02
- Salaschek, M., and Souvignier, E. (2014). Web-based mathematics progress monitoring in second grade. *J. Psychoeduc. Assess.* 32, 710–724. doi: 10.1177/0734282914535719

- Salaschek, M., Zeuch, N., and Souvignier, E. (2014). Mathematics growth trajectories in first grade: cumulative vs. compensatory patterns and the role of number sense. *Learn. Individ. Differ.* 35, 103–112. doi: 10.1016/j.lindif.2014.06.009
- Schurig, M., Jungjohann, J., and Gebhardt, M. (2021). Minimization of a short computer-based test in reading. *Front. Educ.* 6:684595. doi: 10.3389/feeduc.2021.684595
- Seeley, C. L. (2005). “Do the math in your head!” President’s message. Available at: https://www.nctm.org/uploadedFiles/News_and_Calendar/Messages_from_the_President/Archive/Cathy_Seeley/2005_12_mathhead.pdf (Accessed April 26, 2022).
- Selter, C. (2001). Addition and subtraction of three-digit numbers. German elementary children’s success, methods and strategies. *Educ. Stud. Math.* 47, 145–173. doi: 10.1023/A:1014521221809
- Sikora, S., and Voß, S. (2017). Konzeption und Güte curriculumbasierter Messverfahren zur Erfassung der arithmetischen Leistungsentwicklung in den Klassenstufen 3 und 4 [Conception and quality of curriculum-based measurements for the computation performance of primary school students in grade 3 and 4]. *Empirische Sonderpädagogik* 9, 236–257. doi: 10.25656/01:15163
- Soares, N., Evans, T., and Patel, D. R. (2018). Specific learning disability in mathematics: a comprehensive review. *Translational Pediatrics* 7, 48–62. doi: 10.21037/tp.2017.08.03
- Star, J. R., Rittle-Johnson, B., Lynch, K., and Perova, N. (2009). The role of prior knowledge in the development of strategy flexibility: the case of computational estimation. *ZDM* 41, 569–579. doi: 10.1007/s11858-009-0181-9
- Stecker, P. M., Fuchs, L. S., and Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: review of research. *Psychol. Schs.* 42, 795–819. doi: 10.1002/pits.20113
- Stecker, P. M., Fuchs, D., and Fuchs, L. S. (2008). Progress monitoring as essential practice within response to intervention. *Rural Spec. Educ. Q.* 27, 10–17. doi: 10.1177/875687050802700403
- Strathmann, A. M., and Klauer, K. J. (2012). *LVD-M 2–4. Lernverlaufsdiagnostik Mathematik für zweite bis vierte Klassen [Learning Progress Monitoring Mathematics for Second to Fourth Grades]*. Göttingen: Hogrefe.
- Tindal, G. (2013). Curriculum-based measurement: a brief history of nearly everything from the 1970s to the present. *ISRN Educ.* 2013, 1–29. doi: 10.1155/2013/958530
- Torbeyns, J., Ghesquière, P., and Verschaffel, L. (2009). Efficiency and flexibility of indirect addition in the domain of multi-digit subtraction. *Learn. Instr.* 19, 1–12. doi: 10.1016/j.learninstruc.2007.12.002
- Van Der Heyden, A. M., and Burns, M. K. (2005). Using curriculum-based assessment and curriculum-based measurement to guide elementary mathematics instruction: effect on individual and group accountability scores. *Assess. Eff. Interv.* 30, 15–31. doi: 10.1177/073724770503000302
- Varol, F., and Farran, D. (2007). Elementary school students’ mental computation proficiencies. *Early Childhood Educ. J.* 35, 89–94. doi: 10.1007/s10643-007-0173-8
- Verschaffel, L., and Greer, B., and Corte, E. de (2007). “Whole number concepts and operations,” in *Second Handbook of Research on Mathematics Teaching and Learning*, ed. F. K. Lester (Charlotte, NC: Information Age Publishing), 557–628.
- von Oertzen, T., Brandmaier, A. M., and Tsang, S. (2015). Structural equation modeling with Ω nyx. *Struct. Equ. Model. Multidiscip. J.* 22, 148–161. doi: 10.1080/10705511.2014.935842
- Wei, X., Lenz, K. B., and Blackorby, J. (2013). Math growth trajectories of students with disabilities. *Remedial Spec. Educ.* 34, 154–165. doi: 10.1177/0741932512448253
- Wilbert, J. (2014). “‘Instrumente zur Lernverlaufsdiagnostik: Gütekriterien und Auswertungsherausforderungen’ [Tools for learning progress monitoring: quality criteria and challenges with regard to interpretation],” in *Lernverlaufsdiagnostik [Learning progress monitoring]*, eds. M. Hasselhorn, W. Schneider and U. Trautwein. 1st ed (Göttingen: Hogrefe), 281–308.
- Wilbert, J., and Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik [Criteria for analyzing a test measuring learning progress]. *Empirische Sonderpädagogik* 3, 225–242. doi: 10.25656/01:9325
- Winkelmann, H., Heuvel-Panhuizen, M., and Robitzsch, A. (2008). Gender differences in the mathematics achievements of German primary school students: results from a German large-scale study. *ZDM* 40, 601–616. doi: 10.1007/s11858-008-0124-x
- Yarbrough, J. L., Cannon, L., Bergman, S., Kidder-Ashley, P., and McCane-Bowling, S. (2017). Let the data speak: gender differences in math curriculum-based measurement. *J. Psychoeduc. Assess.* 35, 568–580. doi: 10.1177/0734282916649122



OPEN ACCESS

EDITED BY

Jackie Masterson,
University College London,
United Kingdom

REVIEWED BY

Styliani N. Tsesmeli,
University of Patras, Greece
Lénia Carvalhais,
Infante D. Henrique Portucalense
University, Portugal

*CORRESPONDENCE

Michael Schurig
Michael.schurig@tu-dortmund.de

SPECIALTY SECTION

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

RECEIVED 13 May 2022

ACCEPTED 21 November 2022

PUBLISHED 16 December 2022

CITATION

Schurig M, Blumenthal S and
Gebhardt M (2022) Continuous
norming in learning progress
monitoring—An example for a test in
spelling from grade 2–4.
Front. Psychol. 13:943581.
doi: 10.3389/fpsyg.2022.943581

COPYRIGHT

© 2022 Schurig, Blumenthal and
Gebhardt. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Continuous norming in learning progress monitoring—An example for a test in spelling from grade 2–4

Michael Schurig^{1*}, Stefan Blumenthal² and Markus Gebhardt³

¹Faculty of Rehabilitation Sciences, TU Dortmund University, Dortmund, Germany, ²Faculty of Philosophy, Institute for Elementary Education, University of Rostock, Rostock, Germany, ³Faculty of Human Sciences, University of Regensburg, Regensburg, Germany

One of the main goals of the teacher and the school system as a whole is to close learning gaps and support children with difficulties in learning. The identification of those children as well as the monitoring of their progress in learning is crucial for this task. The derivation of comparative standards that can be applied well in practice is a relevant quality criterion in this context. Continuous normalization is particularly useful for progress monitoring tests that can be conducted at different points in time. Areas that were not available in the normalization sample are extrapolated, closing gaps in applicability due to discontinuity. In Germany, teachers participated in a state-funded research project to formatively measure their children's spelling performance in primary school. Data ($N = 3000$) from grade two to four were scaled, linked and translated into comparative values that can be used in classrooms independently from specific times. The tests meet the requirements of item response models and can be transferred well to continuous norms. However, we recommend using the 10th or 20th percentile as cut-off points for educational measures, as the 5th percentile is not discriminating enough.

KEYWORDS

learning progress monitoring, curriculum based measurement, continuous norms, primary school, formative assessment, spelling, learning trajectories

Introduction

In all countries there are children who benefit only slightly or hardly at all from regular instruction in school. International large scale studies (e.g. PISA or PIRLS) show that between 10 and 20% of elementary school children do not acquire the basic skills in reading and mathematics necessary to enter the secondary school (e.g., [Hußmann and Schurig, 2019](#)). Children with barely demonstrable learning growth are often referred to as struggling students or students-at-risk.

Research shows that children's learning development varies and that children learn at different rates depending on the classroom, cognitive prerequisites, motivation, and social environment. In Germany 10–30% of a class show little or no improvement in competencies in Mathematics over a school year, while their classmates show moderate or strong measurable learning growth ([Salaschek et al., 2014](#)). A closer look at the

learning progression in reading and spelling over several years reveals that the gap between high and low achieving students can even widen (DeVries et al., 2018). Lenhard et al. (2017) for example found that reading proficiency levels continue to diverge, especially in the early grades and that the gaps between children's performance remain constant through eighth grade in Germany. Accordingly, Peng et al. (2019) showed that word reading developmental trajectories did not close until a hypothesized performance plateau was reached in the U.S. In a more recent study, Carvalhais et al. (2021) traced the developmental paths at word, sentence, and discourse levels in Portugal. Lower results corresponded to lower academic years between grades 4 and 7 as well as 6 and 9. Discourse-level predictors were identified as the strongest predictors for a written texts' quality in both cohorts while word- and sentence-level predictors only held explanatory power in the younger cohort. This does indicate the need for specific learning difficulties to be signaled in time to appropriately adapt instruction.

Spelling competence is seen as a key qualification in societies. Spelling competence consists of various aspects as punctuation, error sensitivity, correction of spellings and spelling strategies (see KMK, 2005; Jaeuthe et al., 2020). Spelling strategies include both the ability to write words as they are spoken (phonetically) and the consideration of orthographic and morphemic rules. The development of spelling competence is theoretically described as a hierarchically structured competence level model in Germany (see section research questions). Findings in international research show that the acquisition of spelling can be traced back to several components such as L1 (Verhoeven, 2000), linguistic trajectories in word spelling and distinctiveness of cognitive and linguistic trajectories in non-word spelling (Lervåg and Hulme, 2010). Those components therefore have to be addressed in research work.

One of the main goals of the teacher and the school system as a whole is to close learning gaps and support students with difficulties in learning. At the level of the school system, this is labeled as compensatory effects by schools in Germany (e.g., Herrmann et al., 2021). Here, research showed ambiguous results as compensatory effects are found at least as often as so-called Matthew effects where strong students even profit more than students with difficulties in learning (Herrmann et al., 2021). Compensatory effects at the school level are therefore achieved when the school system supports children with learning problems and allows them to catch up with the other students. Current instruction hardly helps these children lagging behind and needs to be changed (Vaughn et al., 2003; Stanat et al., 2017; Fuchs et al., 2021). This leads to the question if there is an international standard for the identification of students-at-risk and students that are in need of individualized education plans. Additionally, the ambitiousness of the support for both groups of students has to be questioned. The short answer is that there are no international quality standards

and often national standards are varying by state or region (Brussino, 2020). National frameworks often remain normative and imprecise (Prince et al., 2018). This leads to the question which economic planning of the funds is efficient and how much individual support is affordable for an education system without withholding resources from students without special needs (Brussino, 2020). In particular, the traditional identification and promotion of special educational needs (SEN) in order to provide more resources to children with learning difficulties is criticized for taking too long, being stigmatizing and not being effective enough (Fuchs et al., 2012).

For children with learning difficulties, support systems with multiple levels of support (MTSS) based on the Response-to-Intervention (RTI) approach have proven to be particularly effective (Fuchs and Fuchs, 2006; Keuning et al., 2017; Arias-Gundín and García Llamazares, 2021) and are now being implemented in more and more countries (Björn et al., 2018). The RTI approach focuses on the learning developments of individual students. It addresses the question to what extent the support works to achieve the learning goal (Fuchs and Fuchs, 2006). To answer this question, students' learning trajectories are monitored and evaluated longitudinally. Thus, for the evaluation of current instructional decision making, several measurements, and information are collected during the learning process since only two pre-post measurements with normed school achievement tests are an insufficient data basis for didactic decisions (Fuchs et al., 2012). Subsequently, decisions about possible adjustments in support are made on the basis of the data collected. Vaughn et al. (2003) see a general paradigm shift away from assessment diagnostics to support diagnostics in the use of data on learning development. Currently, multi-level support systems are implemented in the USA, Finland, the Netherlands and in some regions in Germany (Voß et al., 2016; Björn et al., 2018). In the Netherlands a mandatory participation in the assessment of achievement and achievement development enables data-based adaptive design of instruction on a classroom as well as an individual level. Studies at the school level show positive effects in mathematics and spelling and slightly higher effects for learners with difficulties (van Geel et al., 2016; Keuning et al., 2017). For a comprehensive introduction on the evaluation of intervention programs see Souvignier (2020).

The MTSS usually consists of three levels, which are constructed according to support needs between level 1: "little" to level 3: "need for special education support." Decisions about the level at which students should be supported can be made on the basis of student's scores on screenings, progress monitoring tests and comparison to normalized scores. In summary, this would be called data-based decision-making. Thus, comparison scores are an important benchmark for educational decision-making to determine whether the individual student now needs and can receive more resources (Fuchs et al., 2012).

In this study, comparative scores in spelling were derived using continuous normalization to test the possibility of making gap-free comparisons with a reference group from a federal state Progress Monitoring (PM) platform. In order to understand the ideas for the implementation presented, it is helpful to look at the requirements for PM systems. These are reflected in their quality criteria.

Quality criteria of progress monitoring in spelling

The idea of PM is to provide feedback on the effect of instructional support and interventions over time using repeated short, but reliable standardized tests (Tindal, 2013; Schurig et al., 2021). PM is a form of formative diagnostics that measures and evaluates learning developments and provides direct feedback to teachers and learners (Gebhardt et al., 2021). The aim of PM is to document learning or behavioral development in a precisely formulated area as accurately as possible and necessary, thus enabling teachers to make fact based instructional or educational decisions. The path to the learning goal and the achievement of the defined goal are measured by means of easily manageable short tests as individual learning developments of the students over time (Hosp et al., 2016). This poses multiple substantial and operational challenges. The identification of characteristics and components of the monitored constructs, in the case of this study spelling as an overall competence, but also individual skills for successfully dealing with individual spelling phenomena and an understanding of their interaction, is required as a basis. Spelling development is determined by multiple factors such as cognitive and linguistic components (Lervåg and Hulme, 2010). Mesquita et al. (2020) investigated the spelling abilities of second, third and fourth graders in European Portuguese and addressed the orthographic complexity categories digraph, contextual consistency, position consistency, consonant cluster, stress mark, inconsistency, and silent letter <h>. Differential developmental trajectories per complexity category were found. Kim et al. (2016) found that in Korea learning growth in spelling can be modeled as a function of the orthographic transparency and the differing skill levels of students. Both results indicate that the difficulty of the words to be spelt must be taken into account in the choice of test material. To systematize this difficulty, a review of models of spelling acquisition in German is necessary. In Germany, there is a multitude of models for the development of spelling in primary school. These include models

- Of gradual understanding between the meaningfulness to the lexical order of writing (Brügelmann and Brinkmann, 1994),
- Of the strategies between logographeme (e.g., writing of letters or words from memory) and word spanning (e.g., the orthographically correct composition and choice

of linguistic means through orientation on sentences, paragraphs or whole texts; May, 1990),

- Of phases ranging from proto-alphabetic phonetics to correct spelling with few overgeneralizations (Thom, 2003) or
- Of profiles from an alphabetic/phonologic strategy to an orthographic/grammatical strategy (Reber and Kirch, 2013).

But there are strong intersections in the successive levels (even when connoted as steps, strategies, phases or profiles) of competence, with three levels appearing in all models: (1) Not yet phonetically correct spelling including even scribbled characters or single letters. (2) Phonetically correct spelling with spelling corresponding to pronunciation and (3) orthographically correct spellings with spellings that cannot be explained exclusively by the pronunciation (Jaeuthe et al., 2020). Therefore, a hierarchical structure of spelling competence levels is assumed. But often it remains vague how students are assigned to a developmental step and while common mistakes are attributed to levels students are most often not in longitudinal designs (see Jaeuthe et al., 2020).

The tests have to give a reliable and valid measure of change within students as well as an option to compare growth measures between specific groups of students (Anderson et al., 2017). As with other tests, learning progress monitoring instruments need to address main quality criteria of tests: objectivity, reliability and validity (Good and Jefferson, 1998). However, these criteria must apply not only to data points collected once, but moreover to changes in the data over time. Therefore, homogeneity of the measured constructs over time and sensitivity are—besides the calibration of the tests—also quality criteria for learning progress monitoring tests. Progress monitoring tests must be tested for dimensionality and fairness over time and for different subgroups, so the application of Item-response Theory (IRT) or structural equation modeling is recommended (Wilbert and Linnemann, 2011; Schurig et al., 2021). Criteria that relate to the practical application of tests of learning progress assessment have to be considered too. Social comparison is highly relevant when progress monitoring is used to make statements in relation to the individual reference norm. Accordingly, norms for the change of a competence over time are needed (Hosp et al., 2016; Förster et al., 2017).

For repeated short term measurements of a specific domain, multiple parallel tests and equivalent items are needed to prevent memory effects and different substantial domains from confounding the measures. In parallel test forms, item difficulties within tests differ while the measured domain and overall difficulty are constant between forms of the test (i.e., Embretson, 1996). This way, no additional (possibly varying) variables confound the measures and the difficulty. This can be tested by the analysis of the item parameters as well as the functions of growth. Performance-specific,

potentially non-linear, growth functions over time that can be shown to be as invariant as possible for subgroups are desirable. This leads to the question of the fairness of the PM.

There are different options to view the fairness of a test, though none has been agreed upon generally (e.g., [American Educational Research Association, 2014](#)). The fairness of a test depends on its purpose. One of the main problems in the development of learning progress tests is that each test is supposed to be equally difficult for each observation and that the test has to be fair for all children in the targeted population ([Wilbert and Linnemann, 2011](#); [American Educational Research Association, 2014](#); [Klauer, 2014](#)) including students with SEN. Formative tests must also be comparable for each child across different observations, since performance assessment should relate to individual development over time in the specific dimension being assessed. So, tests have to be equally fair for the same students multiple times.

Test fairness between individuals can be defined as the constancy in difficulty of different groups of test takers within time. For academic progress monitoring items, such as items in spelling, there is the problem that exactly the same items may not be used for each measurement for memory effect reasons. This links to the definition of the test's homogeneity of difficulty. Analysis of differential-item-functioning or measurement invariance can be implied to assess the fairness of the test by group-defining traits (e.g., with or without SEN). If this criterion is met, comparative means may be given to support the usability of the test by giving references to comparable test-takers as well as test-takers for which the test may be too easy or too demanding. This directly addresses the sensitivity of tests.

PM can be constructed for short or longer observation periods. Short, sometimes even weekly, intervals require testing that is as sensitive as possible. This might be used to measure the effect of an intervention in a narrowly defined subarea of a competency or skill ([Hosp et al., 2016](#)). Tests that measure an entire competency (e.g., spelling) may have different tasks from several sub-areas (e.g., different orthographic difficulties such as the number of graphemes or diphthongs). For such tests, measurements with longer time intervals are, for example, monthly, semi-annually, or annually. While tests with shorter observation periods are mainly used for measuring individual learning development, tests with longer measurement periods are also (but not exclusively) used as screenings and as a comparison between students (as in benchmarks).

In addition to measuring the psychometric quality, this also requires an interpretable normalization of the tests with comparisons to age or grade cohorts; the derivation of norms. The challenge is thus: It has to be ensured that the test in question is sensitive enough to detect (eventually small and slow) change within a specific domain ([Kazdin, 2011](#); [Klauer, 2014](#)). This can be achieved by the implementation of appropriate

scoring mechanisms to allow for the comparison of means across time. After a scale is established, mean change, if possible comparisons against national or state-wise percentiles and individual change may be assessed to evaluate the sensitivity of a test across time ([Hasbrouck and Tindal, 2006](#)). But while standardized psychological tests or repeated summative tests are most often taken in equidistant and fixed intervals, tests in PM that are taken in classroom situations will often be taken when convenient. Test times could be omitted or postponed for educational reasons, individual students could be repeatedly absent due to another intervention, or holiday periods could cause gaps in observation and an effect on learning. Therefore, time-independent comparisons are desirable.

For the evaluation of mean change repeated measures analysis of variance might be applied (e.g., [Souvignier et al., 2014](#)). But this is difficult with non-equidistant time intervals. For individual change a function of the observed scores, most often an ordinary least squares regression ([Ardoin et al., 2004](#)), can be computed and evaluated. But this does not address the mean slope (growth) of the comparative sample. For the estimation of mean change latent growth models can be applied, so that latent intercepts and means can be addressed separately ([Förster et al., 2017](#)). Additionally, all analyses have to assume an (often very easy) function of growth, such as a linear or quadratic assumption, which does not account for systematic variations of the population ([Brunn et al., 2022](#)). But the development of students' performance does not necessarily follow linear trajectories ([Strathmann and Klauer, 2010](#); [Salaschek et al., 2014](#); [Mesquita et al., 2020](#)). Furthermore, learning trajectories may differ depending on the study period ([Christ et al., 2010](#)) and baseline level. This could be addressed by large and highly controlled norm samples with multiple points of measurement each.

But how many points of measurement are needed to estimate a (simple linear) slope and make use of the parameters for individual assessment? The [Kratochwill and Levin \(2010\)](#) and [What Works Clearinghouse \(2020\)](#) offer the assessment that five points of data within each evaluated phase are necessary to reach satisfactory coefficient of determination and according error margins. [Christ et al. \(2013\)](#) suggest six to eight points of data. This trait of a test directly refers to usability.

No test is useful if its results are not put to use. Here, the two main approaches are empowering the teachers using the test and simplifying the design of the test ([Deno, 2003](#)). The test's administrators normally are teachers that have received little to none training in the administration and interpretation of diagnostic tools ([van Ophuysen, 2010](#)), stressing the need of a feedback design that takes teachers' understanding, interpretation, and use of data for instructional decision-making into account ([Espin et al., 2017](#)). In addition, the tests have to be designed in practical and usable ways that are easy to teach and time efficient ([Deno, 2003](#)).

No general analytic framework is appropriate in all situations. Interpretations of the results depend on the domain, the difficulty of the test and the intervals between observations (e.g., Hasbrouck and Tindal, 2006). Nevertheless, it is desirable to have comparative values that can be used at any point in the potential study period and that can take into account non-linear developments in several performance levels.

Norming in learning progress monitoring

Fuchs et al. (2021) differentiate in the application of Data Based Decision Making in (a) its use as universal screening at one point of measurement for the performance level, (b) as interpretation of learning development over several weeks, or (c) as interpretation of instructional utility. For each field of application, standards, benchmarks and norms which are useful have to be developed for the individual instruments. Teachers may use norms to have a comparison in addition to individual data to measure learning progress as a basis for their pedagogical decision and basis for the intensity of pedagogical support (Hasbrouck and Tindal, 2006). The standards often refer to curricular settings and the benchmarks are essentially cut-scores that were determined to predict proficient performance at the end of a year (Hosp et al., 2016). It is assumed that for teachers such categories are easier to interpret than continuous scores, if the categories for those are recognized as benchmarks nationwide (Hosp et al., 2016). But this is not always given (see section research questions). Moreover, a fine-grained interpretation of continuous data adds little value to teachers, as this would imply that for every expression of the norm, there is also a routine to support students (e.g., a tier in a RTI). However, the information on continuous variables is lost in this process of categorization (MacCallum et al., 2002). Therefore, in addition to categories, continuous norms for experts in assessment should also be provided. Whether national benchmarks are formed at all and to what extent this is possible in a federal country is an open question. Regional norms are already a step forward if there is no national agreement (e.g., Shinn, 1998).

Depending on the interpretation of the test, depending on the sample and also depending on the scaling of the test, these standardizations and the possible interpretations differ. For educational decision-making, teachers may interpret both the intercept (indicator for level of competence) and the slope (indicator for learning progression) (Hosp et al., 2016). For this purpose, teachers need not only the child's values but also comparative values from standardized school studies (Danielson and Rosenquist, 2014; Förster et al., 2017). Standardized studies are therefore also necessary for the interpretation of individual as well as collective learning goals and trajectories. Norms should be available over at least four measurement time points and should account for children with specific learning difficulties (Förster et al., 2017).

When external criteria are given on how test scores can be interpreted directly there is no need for reference scores on the population. This is because the evaluation of a test score is then conducted in regard to this threshold. However, the vast majority of psychometric tests aim to classify a test result in relation to a reference population (Lenhard et al., 2019). Norm scores represent the distribution of raw scores in a (hopefully the) designated population. The empirical distribution is therefore assessed by a sample that is as representative as necessary. Norms can be expressed in the form of t-values or percentiles. However, since percentiles do not represent a linear transformation of the raw scores, further computation with percentiles may lead to bias. Therefore, the percentiles are usually transformed into norm scores. These may take the form of stanine scores, z-scores, T-scores. The norming of psychometric tests can be defined as setting up population-based reference scores in order to be able to assess the exceptionality of an individual test result (Lenhard et al., 2019).

Traditional norming has limitations on behalf of the sample size, which tend to become rather large due to separated groups and biased percentiles in the extreme values due to floor effects. Additionally extreme values tend to influence percentiles strongly. Discontinuity gaps are often present in norm tables because of the categorical nature of the way time is metrized (Zachary and Gorsuch, 1985). There are no norm values for the time between the time intervals the norming took place in, limiting their usefulness in PM. In the last place, traditional norming is based upon assumptions on normal distributions of the variables.

Continuous norming is based on modeling rather than distributional assumptions. The term continuous norming refers to the statistical modeling of the development of percentiles as a function of the test and further explanatory variables (e.g., age, gender, grade, SEN). The relation between scores and time is computed by the total sample and not by single groups. This way growth can be addressed very accurately by including more parameters (Zachary and Gorsuch, 1985; Lenhard et al., 2018; Voncken et al., 2019). Continuous norms may be calculated by polynomial regression for normally distributed variables (Zachary and Gorsuch, 1985), other assumed distributions (Voncken et al., 2019) or even without distributional assumptions (Lenhard et al., 2018). For a comprehensive summary on continuous norms see Lenhard et al. (2018). A summary of the steps for the derivation of continuous norms without prior assumptions is given in Lenhard et al. (2019). These can be described briefly: (a) Subsamples are created. (b) If a continuous explanatory variable (e.g., age) is used, categorical groups (e.g., age intervals) are generated. (c) For each case position percentiles are identified. (d) For every explanatory variable and every position of each case in a subsample, power and their products are computed. (e) A stepwise regression analysis is done using the powers and their products to predict the empirical raw score. (f) The Taylor

polynomial function is used to predict the raw score based on the explanatory variable(s). (g) The rank is identified with the significant variables from the stepwise regression analysis. Using the identified Taylor polynomial function either norm tables can be generated or norm values can be derived directly based on the measured raw score and age.

Continuous norming is especially relevant for tests where results have to be assessed in regard to age or grade and if the test will be performed at variable times. The potential advantages of continuous norms are therefore the lack of gaps within an age range, a fine age gradation and the extrapolation into ranges that were not available in the sample. Additionally, the required sample size is strongly reduced. In summary, this suggests that continuous norms and standards could be of great importance for the derivation of comparative values in PM procedures. To our knowledge, however, this has not yet been done.

Research questions

The measuring of spelling skills in progress monitoring is usually done by using a robust indicator approach in both primary and secondary schools. The research focuses on identifying appropriate tasks and assessment options. This is because the wide variability of errors in spelling (e.g., capitalization, punctuation, grammar, inflections as well as sequencing) makes it difficult to determine a proxy indicator. In a systematic review by [McMaster and Espin \(2007\)](#) the correct word sequences and the difference of correct minus incorrect word sequences are the most appropriate indicators across all grades. Nevertheless, the number of correctly written words is most commonly used in school practice because this index is reliable and very easy to evaluate ([McMaster and Espin, 2007](#)). Strathmann and Klauer were the first to publish a proposal for a German-language learning progress test to measure spelling competences ([Strathmann et al., 2010](#)). This is a pragmatic dictation test that measures students' transcription skills with the number of correct words as a robust indicator. The spelling of words is assessed according to the categories right or wrong, but it is very easy to design multiple parallel forms of tests that are necessary for progress monitoring. To generate the items, the test authors first created a basic vocabulary ([Strathmann and Klauer, 2010](#)).

In contrast to the USA, there are no benchmarks or standards for PM tests in Germany. In Germany, the federal structure of the education system complicates or even forbids the use of national standards for progress monitoring, as both curricular content, student support and school types differ significantly between the states ([Brussino, 2020](#)). Even basic vocabularies (e.g., the set of words in a language necessary to understand any text in a given language at a given stage of development) differ between states due to dialects (e.g., language varieties). The use of basic vocabulary is an important

didactic approach for the acquisition of spelling skills at school in Germany. Such approaches were already used in the GDR ([Riehme, 1987](#)), and later also in West German states ([Sennlaub, 1985](#); [Naumann, 1987](#)). The reason given for this was the partly limited regularity in German orthography ([Brinkmann and Brügelmann, 2014](#)). Almost all the federal states in Germany have recently developed state-specific basic vocabularies. In total, 1915 words are explicitly listed in the vocabularies of the federal states. However, only 8 words are listed in all. A large proportion of 724 words are listed only in one of the basic vocabularies. For a summary, see [Blumenthal and Blumenthal \(2020\)](#).

Although there are three internet platforms that offer scientifically designed and tested instruments for progress monitoring spelling competencies at the moment ([Blumenthal et al., 2022](#)), the use or application is still rather unknown in practice and not recommended by the state. In Germany, teachers can use the state-funded research project *lernlinie* (learning line¹) to formatively measure their children's spelling performance. The question arises as to whether the available data can be scaled, linked across grades and translated into comparative values that allow individual students to be placed within percentiles. From these considerations, the following research questions were derived:

- Are the test forms IRT scalable and are the reliability values of the person parameters in an acceptable range?
- How strong are the correlations between the person parameters between the points of measurement?
- Is the data sufficient to derive interpretable and meaningful continuous norms?
- What threshold values can be used for the identification of a risk group?
- In a first step, the results of the individual tests are presented and the fit to the Rasch model is demonstrated. In the second step, the norms are formed across the grades using continuous norms.

Methods

Sample and design

The longitudinal study includes 3,000 children from second to fourth grade whose spelling performance was assessed at the beginning of the school year and in the middle of the school year. The scoring of the test was done along the categories of right and wrong. The Internet platform www.lernlinie.de offers free screenings and tests that are appropriate to measure progress over time as a print version under free license for all. However, the use of the platform with the automatic evaluation

¹ www.lernlinie.de

is possible only for teachers of the federal state funding it. This also meant that data protection guidelines of the state were made applicable and relevant background characteristics such as SEN could not be included by the researchers. In the state in question, this information must remain in the schools if guardians have not explicitly released it. This permission was not obtained for the project. In the data analysis, the user data of this platform are evaluated in the spelling tests from second to fourth grade level. Registration and use of the platform are free of charge and voluntary. Teachers can then download the tests as a copy template for a paper-pen test. The teachers enter the students' entries into the database. Then they are analyzed automatically and children's performance are estimated. By now the following normative cut-off points are given: "well below average" for percentile rank <10, "below average" for percentile rank < 25, "average" for percentile rank < 75, "above average" for percentile rank < 90, "well above average" for percentile rank > 90.

The analyses presented here used student data deposited in the database over the 2018/19–2020/21 school years. Tests were administered at 6-month intervals, at the beginning and middle of each school year. The students were distributed among 51 schools from rural or small-town areas. The gender ratio proved to be approximately balanced. General participation was voluntary and schools were free to decide how many and which test dates they participated in. Information on the distribution of students across grade levels can be found in Table 1, the overlap between the participations is given in Table 2.

Instruments

The Reiner test concept records the spelling performance of elementary school children every 6 months at the middle and end of the school year. The test was developed for lernlinie (see Blumenthal, 2022). The test consists of cloze texts (Taylor, 1953) that are dictated by the teacher and for which the target words are to be written down by the children. For the test construction different German textbooks were analyzed. However, the individual spelling phenomena vary in the textbooks, in terms of when they first appear, by up to 3 years (cf. Diehl et al., 2020). Vocabulary and its scope also vary (cf. Voß and Blumenthal, 2020). For the construction of the item pools, the models for reading acquisition in German (Gasteiger-Klicpera and Klicpera, 2005), models for spelling (Reber and Kirch, 2013), and the recommendations of the Standing Conference of the Ministers of Education and Cultural Affairs of the federal states in the Republic of Germany (KMK) for the subject German (KMK, 2005), as well as contents of selected language books for elementary school were used.

The item selection was therefore based on the following criteria:

- Items correspond to the verbal vocabulary of children between six and ten years of age
- Items follow recommendations of education ministries
- Items correspond to the vocabulary of relevant textbooks
- Items represent different levels of difficulty

The recommendations of education ministries means the reference to the intersections of the multiple German basic vocabularies. From several vocabularies, 808 relevant words for elementary school were chosen. From these, a test pool for each grade level was created according to the rules in Table 3. The item pools overlap and due to a linkable multi-matrix design (Mislevy et al., 1992). In multi-matrix designs alternate test forms are created with items from an item pool. For the Reiner Test this resulted in two different forms of the same test (e.g., the same item pool) per grade level. All tests are therefore linked by anchor items. Anchor items are the items that are used in more than one test form to link the results. A total of 275 of the words were used for linking, 98 across grade levels and 177 within grade levels. Attention was paid to a distribution of spelling phenomena to be observed, so that a spread of item difficulties across the anchor items can be assumed (Blumenthal and Blumenthal, 2020). Thus, in grade 1, especially (but not exclusively) phonetic short words were used; in grade 2, mainly phonetic complex or frequent words as well as words with multiple consonants; in grade 3, words with double consonants, compound nouns, the extension *h* or the consonant compounds *ck* or *tz*; in grade 4, words with double vowels, the consonant compound *chs*, adjectives ending in *-ig* or foreign words.

The tests were piloted and a main study with $N = 4091$ children in 192 first to fourth grades and 24 schools showed fit to a unidimensional Rasch model (Voß and Sikora, 2017). The levels of difficulty were chosen in accordance to Embretson (1996) in order to cover a full range of abilities and thus to enable the location of the person parameters against the background of the differing item difficulties. Words were chosen by the item parameters (Blumenthal and Blumenthal, 2020) to represent easy, medium and difficult words. From a psychometric point of view, 732 words could be identified as suitable for assessing spelling competencies from the grade level 2–4.

The formal design of the spelling tests was guided by economic and pragmatic factors that an inclusive school setting entails (Hosp et al., 2016). For example, they were to be feasible as group procedures in a class setting, the test was not to last longer than 15 min and they were to include tasks that were close to instruction, such as simple word dictations with word counts that depended on the grade level and on whether the test was taken at the beginning or in the middle of the school year (grade 2: first test 24 items and second test 36 items, grade 3: first test 36 items and second test 48 items, grade 4: first test 48 items and second test 60 items). All target words were embedded in narrative texts around the identification figure (a pig named Reiner) that were appealing

TABLE 1 Students by grade and gender.

	Grade and test within the grade					
	Grade 2		Grade 3		Grade 4	
	t1	t2	t1	t2	t1	t2
Boys	604	324	688	282	561	238
Girls	585	312	589	232	480	211
Total	1,189	636	1,277	514	1,041	449

TABLE 2 Overlap between grades.

	Number of participations						Total
	6 times	5 times	4 times	3 times	2 times	1 time	
Boys	8	39	74	201	296	963	1,581
Girls	10	34	51	195	261	868	1,419
Total	18	73	125	396	557	1,831	3,000

TABLE 3 Structure of the reiner tests.

Grade 2	Grade 3	Grade 4
1. Phonetic words with 3-4 graphemes	1. Common words	1. Common words
2. Phonetic words with 5-8 graphemes	2. Phonetic words with 2-3 graphemes	2. Phonetic words without restriction of the number of graphemes
3. Special/difficult words (diphthongs, umlauts, words with v)	3. Phonetic words with 4-7 graphemes	3. Words with diphthongs au, ei
4. Words with double consonant	4. Words with [ie]	4. Words with [ie]
5. Words with [ck]	5. Words with [ß]	5. Words with [ß]
	6. Words with [qu]	6. Words with [x]
	7. Words with [ck]	7. Words with [tz]
	8. Words with [v] at the beginning of the word	8. Words with [ck]
	9. Words with umlaut [ä], [ö], [ü]	9. Words with extensions [üh], [ieh], etc.
	10. Words with consonant doubling ll, tt, nn, mm	10. Words with consonant doubling ll, tt, ff, ss
	11. Words with multiple consonants (e.g. nst)	11. Words ending in -ig, -lich, -ung, -heit, -keit
	12. Words with stretching h	12. Words with prefixes be-, ge-, ent-, ver-, vor-
	13. Word combinations	13. Words ending with [chs], [ks]
		14. Words with a double vowel
		15. Words with [qu]
		16. Words with stretching h
		17. Words with pronoun hardening
		18. Words with [v] at the beginning of a word
		19. Words with vowel derivatives to umlaut [a-ä], [u-ü]
		20. Word combinations
		21. Foreign word
Examples (Grade 2 Test 2)	Examples (Grade 3 Test 2)	Examples (Grade 4 Test 2)
Wo [where]	dem [the (dative)] Lied [Song] springst [you are jumping]	Glück [Luck]
vom [from]	Blätter [Leaves] Geburtstag [Birthday]	gießen [we are casting]
Euro [Euro]		Frühling [Spring]
Lasso [Lasso]		ängstlich [anxious]
Körper [Body]		unglaublich [unbelievable]

to children in order to a) increase motivation to complete the tests and b) generate content contexts for semantically ambiguous words. It is assumed that spellers with difficulties might use context clues to their advantage (Taylor, 1953; Ehri, 2005).

The complexity of the texts was determined using the LIX index (Lenhard and Lenhard, 2014). The LIX index accounts for surface features of a text (number of words, word length, proportion of long words, and sentence length) and thus forms an indicator for assessing its difficulty or ease. The LIX is the sum of the average sentence length of a text and the percentage of long words (more than six letters). Care was taken to ensure that the LIX values for the text templates were below 40 and could thus be assumed to be suitable for children and adolescents.

Initial analysis of the psychometric adequacy of the developed tests revealed high reliabilities in the range between $.90 \leq \alpha \leq 0.96$. Correlations with convergent procedures (spelling test Hamburger Schreib-Probe 1-10; May et al., 2019) vary between $r = 0.69$ ($N = 56$) and $r = 0.82$ ($N = 177$) and testify to the validity of the instruments. Further evidence of the psychometric quality of the tests was determined in the present study. In Table 4 the descriptive values as well as the accuracy of the tests within the grades are given.

Results

The basic psychometric criteria were analyzed by the application of IRT analysis with TAM (Robitzsch et al., 2021) in R (R Core Team, 2021). One-parameter logistic models with marginal maximum likelihood estimators were applied. The random effect models were done with lme4 (Bates et al., 2015) and the visualization was done with GAMLj (Gallucci, 2019) in jamovi (The jamovi project, 2022).

There are different approaches to the computation of continuous norms. Parametric approaches are making assumptions on the distributional shape of the raw scores (e.g., R Package GAMLSS; Rigby and Stasinopoulos, 2005), non-parametric regression based approaches and semi-parametric approaches address the norms as latent variables (Lenhard et al., 2018; R Package cNORM). The approach within cNORM offers the beneficial characteristic that it does not require any distributional assumptions. Therefore, in most use cases the data can be modeled more precisely than with parametric methods (Lenhard et al., 2019). This is particularly true for small samples as small as < 100 and skewed distributions. The data as well as the code is deposited as an open-access OSF project (Schurig et al., 2021²).

² <https://osf.io/vg2r7/>

Scaling

To assess the fit of the models measures of reliability (EAP Reliability), a measure of local independence (the average of absolute values of the adjusted Yen's Q3 statistic; Yen, 1984, see Robitzsch et al., 2021) as well as mean Outfit and Infit values are given (Table 5; Wright and Masters, 1982). The necessary criteria were met within each grade with reliabilities exceeding 0.8, the mean Q3 statistics approximating 0 and the mean item fit values approximating 1. It can be assumed that the usage of sum scores is defensible (Rost, 2004).

To address the fairness of the test effects of differential item functioning (DIF) between gender groups were analyzed by using Raju's Area method (see Wright and Oshima, 2015) implemented in Snow IRT (Seol, 2022). For this method effects sizes were introduced by Wright and Oshima (2015) with cut-off values between < 1 for neglectable effects and > 1.5 for large effects of DIF (Magis et al., 2010).

In Grade 2 time 1 no effects > 1 were observed so that a negligible DIF can be assumed ($M_{abs} = 0.37$, $SD = 0.23$) and in 2.2 one moderate and one large effect were observed ($M_{abs} = 0.48$, $SD = 0.41$). In Grade 3 time 1 one item showed a large effect ($M_{abs} = 0.40$, $SD = 0.44$) and in Grade 3 time 2 six items showed moderate and one item showed a large effect ($M_{abs} = 0.54$, $SD = 0.38$). In grade 4 time 1, one item showed a large effect ($M_{abs} = 0.49$, $SD = 0.31$). In Grade 4 time 2 moderate effects were observed in nine and large effects were observed in six items ($M_{abs} = 0.74$, $SD = 0.60$). However, the effects do not have a clear direction which could be interpreted, so that one can assume random and therefore ignorable DIF effects. In the next step the distributions of the sum scores are given. As can be seen in Figure 1 the distribution of the percentage of accuracy is becoming more skewed toward the higher grades. On the y-axis the density is given due to different sample sizes.

When the measures between the points of measurement are correlated, the effects (Spearman rank correlation coefficients; Table 6) range from $r_s = 0.51$ to 0.85. Roughly speaking, the effects are higher the closer the data points lie to each other in time.

Random effects modeling

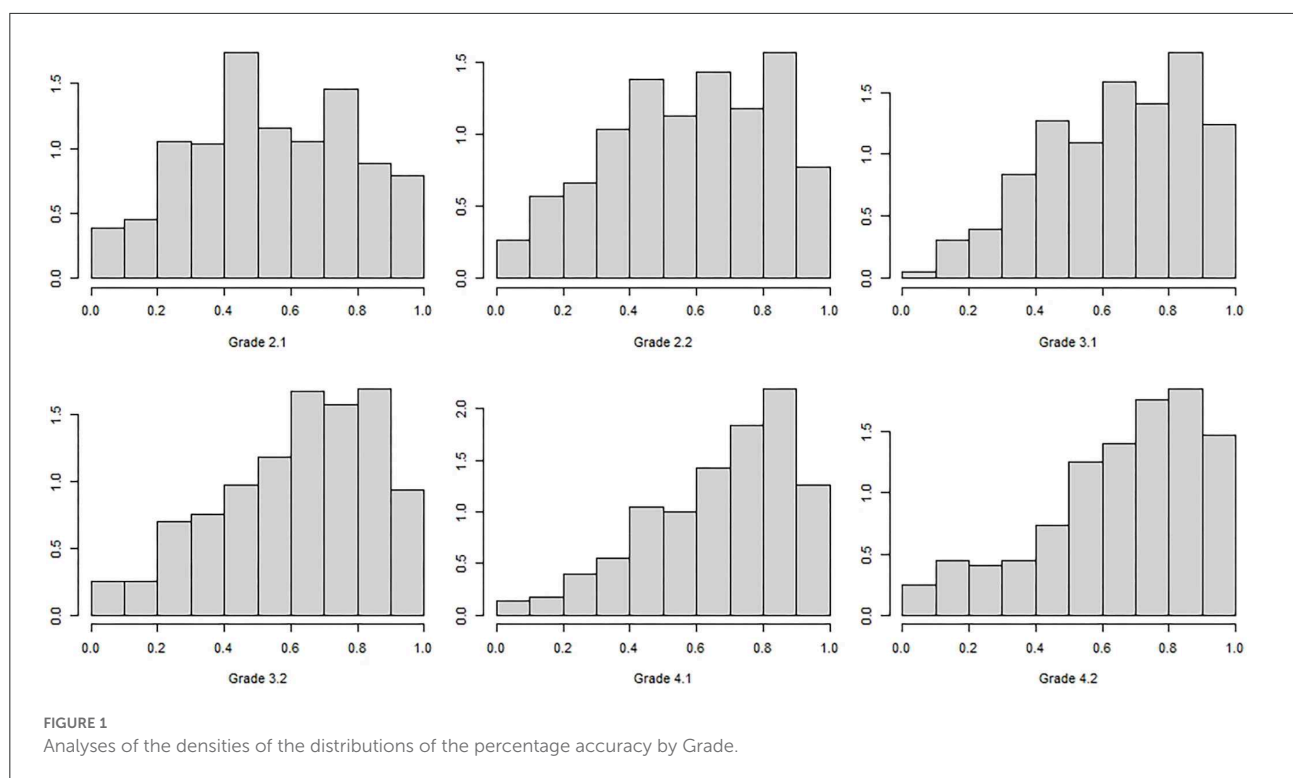
In a last step individual growth effects were addressed with a linear mixed model with the scores as random effects within persons and time as a factor (Gallucci, 2019). Here two models were taken into consideration. In the first place (Model 1; Figure 2) the percentage accuracy was analyzed as a random effect. In both models the number of included cases is $n_{id} = 2981$. This is the number of students with at least two successive points of measurement and the number of observations within the cases is $n_{obs} = 5087$. In the second place (Model 2; Figure 2) the sum of the solved items was analyzed. In Model 1 the stability of the difficulty of the test is addressed. In Model 2 the individual

TABLE 4 Descriptive values of the reiner tests.

Grade	Timepoint	N	Mean	Median	SE	SD	Min.	Max.	Missing	Perc. Acc.	SD
2	1	1,189	13.1	13	0.17	5.87	0	24	0	0.55	0.24
2	2	636	20.8	21	0.35	8.71	0	36	0	0.58	0.24
3	1	1,277	23.2	24	0.22	7.99	0	36	0	0.64	0.22
3	2	514	29.7	31	0.5	11.3	0	48	0	0.62	0.24
4	1	1,041	32.2	34	0.32	10.4	0	48	0	0.67	0.22
4	2	449	39.8	43	0.67	14.3	3	60	0	0.66	0.24

TABLE 5 Fit statistics of the used measures.

Grade	<i>t</i>	<i>n</i>	# items	EAP Rel.	MADaQ3	<i>M</i> Outfit	<i>SD</i> Outfit	<i>M</i> Infit	<i>SD</i> Infit
2	1	1,189	24	0.88	0.084	0.99	0.14	1.00	0.06
2	2	636	36	0.92	0.048	1.01	0.21	1.00	0.10
3	1	1,277	36	0.90	0.051	0.99	0.19	1.00	0.11
3	2	514	48	0.93	0.046	1.04	0.37	1.00	0.12
4	1	1,041	48	0.92	0.043	0.99	0.19	1.00	0.10
4	2	449	60	0.94	0.048	1.01	0.32	1.00	0.12



growth in dependence on the increased length of the test (24–60 items) and the increased difficulty of the items (see Instruments) is in the center of interest. The Pearson correlation between the sum scores of all observations and the percentage solved is $r = 0.822$ ($p < 0.001$). Figure 2 (Model 1) indicates that there is

no significant floor effect for the difficulty of the tests. The figure of Model 2 shows that there is a compression at the ceiling of the test (the maximum number of items) but that the test also covers low performance and its development, especially from grade 3 onwards.

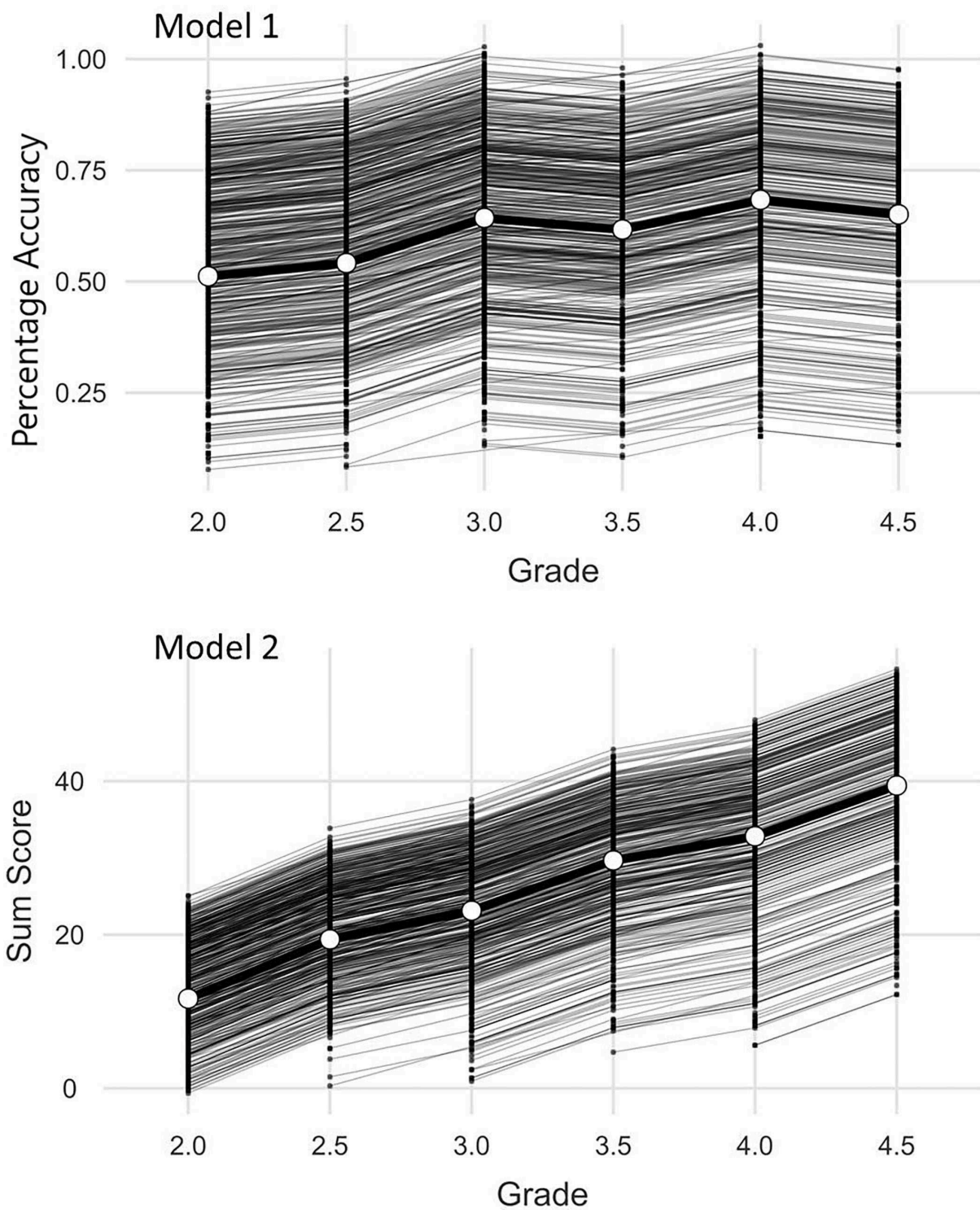


FIGURE 2
Analyses of random effects models.

For both Models fixed effect omnibus tests (Wald) showed significant main effects (Model 1: $F(5, 2815) = 151, p < 0.001$;

Model 2: $F(5, 2990) = 1365, p < 0.001$). The conditional R^2 (variance explanation of the model) in Model 1 is 0.79 and in

Model 2 it is 0.86. The marginal R^2 (variance explanation by time alone) is 0.07 in Model 1 and 0.47 in Model 2. The intra class correlation (ICC) of the random component (Student ID) is 0.78 in Model 1 and 0.74 in Model 2. This indicates an expected high variance explanation that can be attributed to the student's proficiency. The low marginal R^2 of Model 1 shows that the difficulty of the test does change significantly but only slightly on behalf of the effect sizes over time which is desirable for the successful linking of the measured values. The fixed effects in Model 2 are larger due to the rising ceiling of the test (Table 7).

In summary, it can be stated that the available data are suitable to a sufficient extent to derive standard values. The relevant question for the linked distributions is whether it is possible to cover a sufficiently broad range of abilities to derive percentiles of interest. Since the aim of these percentiles is to identify students who have difficulties in learning, the relevant question is which percentile is chosen to derive learning difficulties. This can be deduced directly from the assumed volume of a tier of the RTI system implemented nationally or regionally. "In a well-designed RTI system, primary prevention should be effective and sufficient for about 80% of the student population" (National Center on Response to Intervention, 2010). In the model project, it was analogously stated that it can be assumed that ~20% of the students are supposed to receive second tier support (secondary prevention) and up to 5% of the students are supposed to receive intensive individual support (Voß et al., 2016). Taking into account the level of error, progress monitoring, which is used for screening and in addition to possible assessment diagnostics to decide on a support tier, should be selective, especially in the lowest quartile for the second tier (25%) and the lowest percentile for the third tier (10%).

Normalization

Since multiple regression is used to obtain a model which allows for the estimation of normal values the first step is to identify this model. cNorm utilizes the best regression subset approach to do so (James et al., 2013). The approach returns a regression model, which describes the given norm sample as well as possible with a minimal number of predictors (Gary et al., 2021). These are the explanatory variables (e.g., age or grade), the powers as well as the interactions of person location on the spectrum and explanatory variable. After a model is established, a numerical approximation (not an exact calculation) of the norm score in question can be deduced. For the necessary ranking of the person scores in the grade groups the default procedure was chosen. The degree of the polynomial of the regression function was chosen to be quartic. For the normalization all cases ($N = 3000$) were included even though the sample sizes varied strongly. $n = 1831$ students

only took part once. $n = 557$ took part two times, $n = 396$ three times and so forth (see Table 2 in section sample and design). $n = 1419$ girls and $n = 1581$ boys took part at all.

For the model validation an adjusted R^2 value can be used. This is the representation of the approximation of the polynomial on the person score, the estimated norm score and in this case the grade variable. The modeling procedure of the Reiner scores from Grade 2 time 1 to grade Grade 4 time 2 reached an adjusted $R^2 = 0.991$ (which also is the stopping criterion of the Taylor function in cNorm; Lenhard et al., 2018) with five terms and an intercept. The number of terms was cross validated (20% validation sample; see Gary et al., 2021), repeated ten times and with up to ten terms. No substantial improvement in model fit could be achieved by adding more terms.

Three powers and the related interactions were needed to fit a sufficient model. The root mean square error (RMSE), deduced from the difference between the predicted scores and the manifest scores, reaches $RMSE = 0.0224$. When taking into account that the person score in question is a percentage with possible values between 0 and 1 the error is justifiable. In Figure 3 it can be seen that the fit is worse in the area of extreme values especially in the higher grades but in most cases the fitted scores are approximating the observed scores well.

The observed and predicted percentile curves are given in Figure 4. PR stands for percentile rank and the following number the percentile. For example, PR50 describes the 50th percentile. It can be seen that the changes of the test designs (more and more difficult items) are reflected in the observed normal values. A high proximity of the percentile curves to each other indicates poor separability between these curves as can be seen above the 75% percentile. The lower ranks are clearly separable though.

To check which percentiles can be loaded in terms of content, it is possible to inspect the confidence intervals of the norm scores (here T-values) within the measurement time points. This is relevant for the research question on possible threshold values for students-at-risk. However, since these are estimated from the complete population, they reach roughly 10% with a 90% confidence interval, regardless of rank and when controlling for regression to the mean, given the smallest observed reliability of 0.88 [for details on the estimation of the C.I. see Lenhard et al. (2018); see Supplementary Table 1].

With reference to the underlying test, however, it is desirable to achieve a high discriminatory power for the ability range that the test is intended to cover in particular. In the case of the present test, this corresponds in particular to the threshold value necessary to separate the 10th and 25th percentiles or (very) roughly t values between 30 to 40 points. In terms of raw scores, this corresponds to a percentage of solved items between 20 and 30% in the 10th percentile and between 40 and 50% in the 25th percentile. The test is therefore easy enough to include information to separate between ranks in the relevant ability domain.

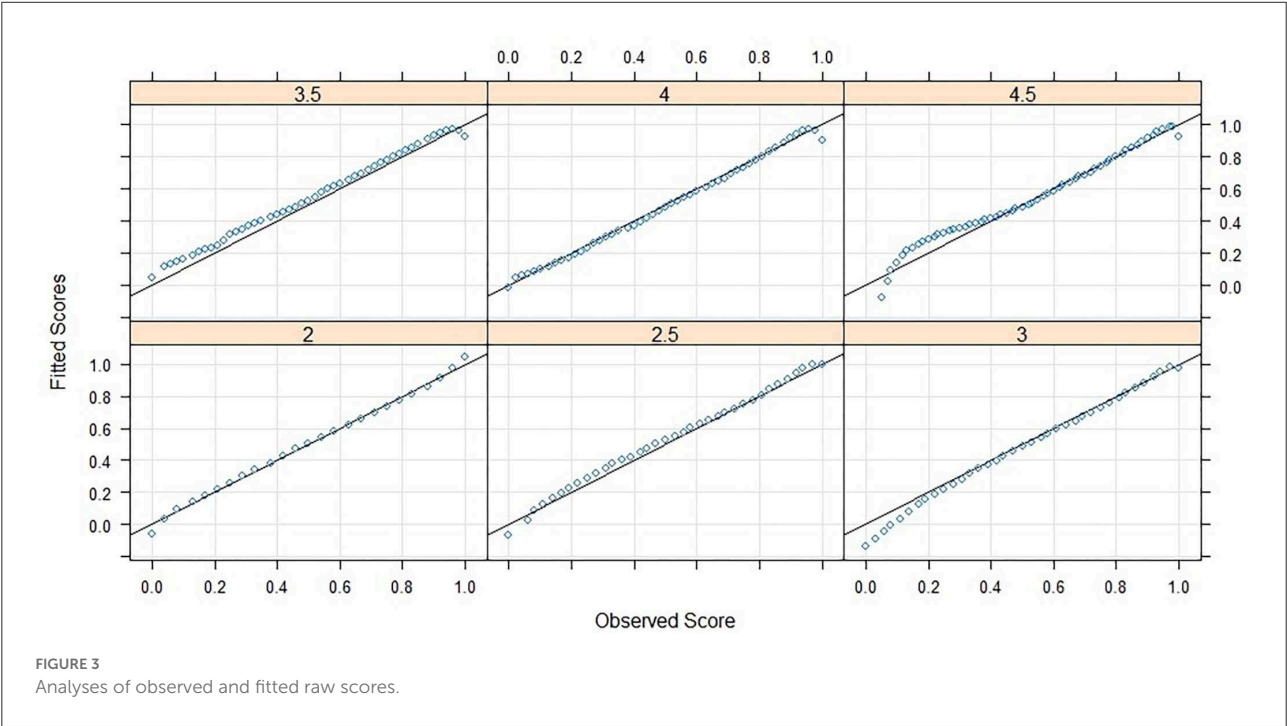
TABLE 6 Spearman correlations between the points of measurement.

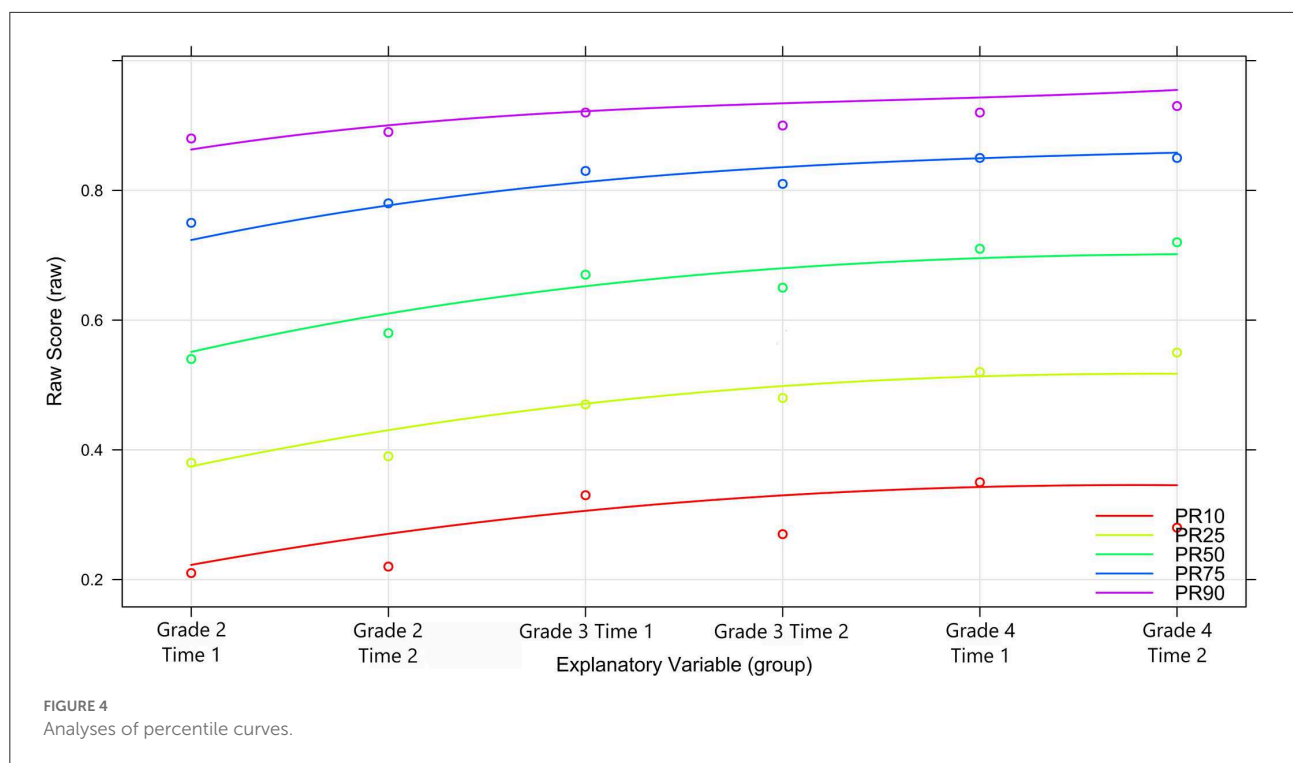
		Grade 2		Grade 3		Grade 4	
		t1	t2	t1	t2	t1	t2
Grade 2	t1	1					
	t2	0.74	1				
Grade 3	t1	0.67	0.83	1			
	t2	0.68	0.83	0.79	1		
Grade 4	t1	0.58	0.74	0.67	0.79	1	
	t2	0.51	0.75	0.72	0.85	0.84	1

Only pairwise complete observations were used. All Correlations are significant on a $p < 0.01$ level.

TABLE 7 Fixed effects parameter estimates.

Model 1 (Percentages)					Model 2 (Sum scores)			
Names	Est.	95% confidence interval		<i>p</i>	Est.	95% confidence interval		<i>p</i>
		Lower	Upper			Lower	Upper	
Intercept	0.61	0.60	0.62	< 0.001	26.02	25.69	26.36	< 0.001
2.5–2.0	0.03	0.02	0.04	< 0.001	7.71	7.16	8.26	< 0.001
3.0–2.0	0.13	0.12	0.14	< 0.001	11.45	10.95	11.95	< 0.001
3.5–2.0	0.11	0.09	0.12	< 0.001	17.99	17.34	18.64	< 0.001
4.0–2.0	0.17	0.16	0.19	< 0.001	21.14	20.55	21.73	< 0.001
4.5–2.0	0.14	0.12	0.16	< 0.001	27.75	27.01	28.49	< 0.001





Discussion

The tests are clearly scalable and reliable and the person parameters correlate strongly across the times of measurement. The results show that achievement gaps between students in our study generally increase over the years. This finding is in line with Herrmann et al. (2021). This is particularly critical given that the teachers involved in our study received the results from the tests in the sense of a formative assessment. Thus, even despite this information, the students with the most difficulties did not succeed in catching up with the rest of the class. But we do not know to what extent the information was used. The interpretation must take into account that the study was conducted in inclusive schools. Thus, this result is also in line with the research that children with special educational needs differ significantly from the performance of normal students and also fail to catch up with this performance by the end of school (Gebhardt et al., 2015). Even with a comprehensively designed system for high-quality instruction for all students and effective support for at-risk children, it may not be possible to adequately address the needs of all children (Voß et al., 2016). The results of the long-term study of the inclusion model in Rügen (Blumenthal et al., 2019) show that prevalence of special needs has been significantly reduced. However, there is a not inconsiderable proportion of students with extensive difficulties at school for whom long term support must also be offered. This is not only a regional

phenomenon, but is also evident in the international context (Fuchs et al., 2014, 2017). Research indicates that 5 to 10% of the student population requires intensive intervention in terms of special education support (O'Connor and Fuchs, 2013).

Normalized scores could be readily derived by the applied procedure. But the question on possible thresholds for the identification of students-at-risk has to be answered in regard to error margins of percentiles. The statistical results show that a representation of the fifth percentile range is associated with too large errors in this study. Therefore, such a cut-off is rather inappropriate for extensive educational decisions based upon the test in question. However, it seems appropriate to consider percentile 10 as the smallest cut-off line. We understand MTSS as a tiered system, in a pragmatic approach. It would be nice to determine the exact level of all learners at all points in time, but that doesn't work without a lot of effort and (very long) tests. Ergo: We stick to an indicator that roughly signals to us that something is wrong and then we take a closer look to initiate and optimize support processes. Ultimately, setting a threshold for student achievement is a normative decision. It could be shown that the present test can support this in the range of the 10th percentile. However, whether this is appropriate or whether individual consideration should be given to the 25th percentile is also a decision that must take into account the performance of a school system. A Smart RTI System as proposed by Fuchs et al. (2012) does not rely

on error-free measurement on every level and at every point of measurement. In general, it can be assumed that norm statements for teachers should rather refer to coarser categories (percentile limits 10, 25, 50, 75, 90). These serve data-based decisions in terms of level assignment in an MTSS. Finer gradations (as can readily be found in continuous norms) are associated with higher probabilities of error and add little value here. The classification of students between these thresholds over time, without the need for a fixed time interval for testing, can thus make it possible to classify them in MTSS systems with sufficient certainty for this purpose. It must be clear, however, that a single measurement is not sufficient to map a learning process and that this single measurement cannot be seen as a “substitute” for a status diagnosis (Christ et al., 2013). Finer norms would also suggest that the school also has concrete measures and responsibilities ready for all gradations. The categorization of norms is accompanied by a considerable loss of information, which can have a significant impact, e.g., when a child’s performance falls very close to the borderline between percentiles. In this respect, combined information in the sense of positioning student performance in a percentile band with additional specification of a confidence interval is important. Within the framework of scientific research, fine norm gradations can also be processed and taken into account accordingly by means of different analysis methods. Here, a loss of information through data categorization would be detrimental. The most important question is how to design funding and resources so that children with more needs get more effective support without being stigmatized (Meijer and Watkins, 2019). The application of the norms in screening help for the application of a multilevel support system to make an important basis for the pedagogical decisions.

The following limitations of the current study must be considered. The sample used is selective and insufficiently controlled to determine the effects of compensatory measures on a school or even a classroom level. For example, the quality of the instruction at the classroom level could not be determined. Furthermore, the data collected is sufficient to model latent trajectories on a growth level but is not sufficient to model individual learning trajectories due to irregular participation. The Reiner test tends to have ceiling effects because the number of items per test is limited and only those items that correspond to the grade level spelling instruction were selected. However, this is negligible for screening purposes in the lower percentiles. Also it has to be mentioned that the test is designed as a group test with a dictation. Thus, the test is not designed to be administered individually. Another limitation of the results is the lack of comparison with an external characteristic on the basis of which the specificity and sensitivity of the results could be demonstrated over time. Differential results of Verhoeven (2000) or Lervåg and Hulme (2010) could also not be replicated due to the lack of background characteristics in this sample. For

a better generalizability of the results, a sample with a higher rate of control is needed. Lastly the grade bracket of this study did not include data from first grade due to changes in the test. This shortcoming has to be addressed.

Developing screenings and progress monitoring instruments to identify children with learning difficulties is important, but not easy (Fuchs et al., 2021). The tests must be both easy to use and to interpret by teachers as well as psychometrically tested and reliable (Schurig et al., 2021). The Reiner test was constructed according to the needs of teachers and the regulations of the school system and was also able to demonstrate psychometrically appropriate goodness. A level of difficulty was selected for each grade level so that the test met the requirements of the grade level. Those grade levels aligned well across time but one should interpret the course over all 4 years only cautiously. Overall, there is considerable variation between the children, which increases rather than decreases over the years. While the test measures a more restricted test range in the first years, the test range becomes larger over the years with further requirements. Since there are fewer but still some easy items in the higher grade level test, the Reiner tests are also very sensitive to the lower percentiles.

Why are there no visible compensation effects of the tests? It has to be stated that formative assessment is still not widely used and teacher professionalism is expandable. The Reiner test concept is already in use in the inclusive region of Rügen and has proven itself as a standardized instrument with comparable norms. This makes it one of the three instruments used in Germany (Blumenthal et al., 2022). It offers a longitudinal screening with clear curricular references as well as a qualitative diagnostic that is linked to proposed pedagogical intervention. This outlines clear support structures. But in the final step, schools and sometimes even teachers in Germany decide for themselves how to deal with such offers due to the high degree of autonomy. This also includes the textbooks used and the question of the closeness to the textbooks in the design of the instruction at a classroom level. However, there is a lack of implementation in the system, training, etc. The next step for Reiner will be to examine the extent to which testing can be implemented more regularly per class and whether and where test results can be integrated into everyday school life and the associated support in learning.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: OSF <https://osf.io/vg2r7/>, doi: 10.17605/OSF.IO/VG2R7.

Ethics statement

The studies involving human participants were reviewed and approved by the Ministry for Education and Child Day Promotion, Germany/Mecklenburg-Western Pomerania. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

MS did the writing and the analyses. SB provided the data and co-wrote the article. MG co-wrote the article. All authors contributed to the article and approved the submitted version.

Funding

We acknowledge financial support by Deutsche Forschungsgemeinschaft and TU Dortmund within the funding program Open Access Publishing.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Anderson, D., Kahn, J. D., and Tindal, G. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Appl. Measur. Educ.* 30, 163–177. doi: 10.1080/08957347.2017.1316277
- Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychol. Rev.* 33, 218–233. doi: 10.1080/02796015.2004.12086244
- Arias-Gundín, O., and García Llamazares, A. (2021). Efficacy of the RtI model in the treatment of reading learning disabilities. *Educ. Sci.* 11:209. doi: 10.3390/educsci11050209
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Björn, P. M., Aro, M., Koponen, T., Fuchs, L. S., and Fuchs, D. (2018). Response-to-intervention in Finland and the United States: mathematics learning support as an example. *Front. Psychol.* 9:800. doi: 10.3389/fpsyg.2018.00800
- Blumenthal, S. (2022). Lernlinie Diagnose- und Fördermaterialien [Learning Line - Diagnostic and Support Materials]. Available online at: https://www.lernfortschrittsdokumentation-mv.de/_lernlinie/index.htm
- Blumenthal, S., and Blumenthal, Y. (2020). Brav ist schwer, Vogel ist leicht. Eine Analyse geläufiger Mindestwortschätze im Deutschunterricht ["Brav" is difficult, "Vogel" is easy – An analysis of German common basic vocabularies in elementary school]. *Empirische Sonderpädagogik* 4, 279–294. doi: 10.25656/01:21612
- Blumenthal, S., Gebhardt, M., Förster, N., and Souvignier, E. (2022). Internetplattformen zur Diagnostik von Lernverläufen von Schülerinnen und Schülern in Deutschland. Ein Vergleich der Plattformen Lernlinie, Levumi und quop [Internet platforms for the diagnosis of students' learning trajectories in Germany. A comparison of the platforms Lernlinie, Levumi and quop]. *Zeitschrift für Heilpädagogik* 73, 153–167.
- Blumenthal, Y., Voß, S., Sikora, S., and Hartke, B. (2019). "Selected findings of the first large-scale implementation of Response to Intervention in Germany" in *Inclusive Mathematics Education. State-of-the-Art Research from Brazil and Germany*, eds D. Kolloche, R. Marcone, M. Knigge, M. G. Pentead, and O. Skovsmose (New York, NY: Springer), 123–145. doi: 10.1007/978-3-030-11518-0_10
- Brinkmann, E., and Brügelmann, H. (2014). "Konzeptionelle Grundlagen und methodische Hilfen für den Rechtschreibunterricht: Schreiben lernen, Schreiblernmethoden und Rechtschreiben lernen in der Grundschule." Available online at: https://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/faecher/sprachen/deutsch/schreiben_rechtschreiben/Konzeptionelle_Grundlagen_Rechtschreibunterricht.pdf (accessed July 19, 2022).
- Brügelmann, H., and Brinkmann, E. (1994). "Stufen des Schriftspracherwerbs und Ansätze zu seiner Förderung" in *Wie wir Recht Schreiben Lernen. 10 Jahre Kinder auf dem Weg zur Schrift*, eds H. Brügelmann and S. Richter (Lengwil: Libelle), 44–52.
- Brunn, G., Freise, F., and Doebl, P. (2022). Modeling a smooth course of learning and testing individual deviations from a global course. *J. Educ. Res.* 14. doi: 10.31244/jero.2022.01.05
- Brussino, O. (2020). *Mapping Policy Approaches and Practices for the Inclusion of Students With Special Education needs: OECD Education Working Papers*.
- Carvalho, L., Limpo, T., and Pereira, L. Á. (2021). The contribution of word-, sentence-, and discourse-level abilities on writing performance: a 3-year longitudinal study. *Front. Psychol.* 12:668139. doi: 10.3389/fpsyg.2021.668139
- Christ, T. J., Silberglitt, B., Yeo, S., and Cornier, D. (2010). Curriculum-based measurement of oral reading: an evaluation of growth rates and seasonal effects among students served in general and special education. *School Psychol. Rev.* 39, 447–462. doi: 10.1080/02796015.2010.12087765
- Christ, T. J., Zopluoglu, C., Monaghan, B. D., and van Norman, E. R. (2013). Curriculum-based measurement of oral reading: multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *J. School Psychol.* 51, 19–57. doi: 10.1016/j.jsp.2012.11.001
- Danielson, L., and Rosenquist, C. (2014). Introduction to the TEC special issue on data-based individualization. *TEACHING Except. Child.* 46, 6–12. doi: 10.1177/0040059914522965
- Deno, S. L. (2003). Developments in curriculum-based measurement. *J. Spec. Educ.* 37, 184–192. doi: 10.1177/00224669030370030801

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.943581/full#supplementary-material>

- DeVries, J., Rathmann, K., and Gebhardt, M. (2018). How does social behavior relate to both grades and achievement scores? *Front. Psychol.* 9:857. doi: 10.3389/fpsyg.2018.00857
- Diehl, K., Hartke, B., and Mahlau, K. (2020). *Inklusionsorientierter Deutschunterricht. Handlungsmöglichkeiten Schulische Inklusion*. Stuttgart: Kohlhammer.
- Ehri, L. C. (2005). "Development of Sight Word Reading: Phases and Findings," in *The Science of Reading: A Handbook*, eds M. Snowling and C. Hulme (Malden, Mass. Blackwell), 135–154. doi: 10.1002/9780470757642.ch8
- Embretson, S. E. (1996). The new rules of measurement. *Psychol. Assess.* 8, 341–349. doi: 10.1037/1040-3590.8.4.341
- Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., and Rooij, M., de (2017). Data-based decision-making: developing a method for capturing teachers' understanding of CBM graphs. *Learn. Disabil. Res. Pract.* 32, 8–21. doi: 10.1111/ldrp.12123
- Förster, N., Kuhn, J.-T., and Souvignier, E. (2017). Normierung von verfahren zur lernverlaufsdiagnostik [Establishing measures of growth in learning progress assessment]. *Empirische Sonderpädagogik* 116–122. doi: 10.25656/01:14998
- Fuchs, D., and Fuchs, L. S. (2006). Introduction to response to intervention: what, why, and how valid is it? *Read. Res. Quart.* 41, 93–99. doi: 10.1598/RRQ.41.1.4
- Fuchs, D., Fuchs, L. S., and Compton, D. L. (2012). Smart RTI: a next-generation approach to multilevel prevention. *Except. Children* 78, 263–279. doi: 10.1177/001440291207800301
- Fuchs, D., Fuchs, L. S., and Vaughn, S. (2014). What is intensive instruction and why is it important? *TEACHING Except. Child.* 46, 13–18. doi: 10.1177/0040059914522966
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Stecker, P. M. (2021). Bringing data-based individualization to scale: a call for the next-generation technology of teacher supports. *J. Learn. Disabil.* 54, 319–333. doi: 10.1177/0022219420950654
- Fuchs, L. S., Fuchs, D., and Malone, A. S. (2017). The taxonomy of intervention intensity. *Teach. Except. Child.* 50, 35–43. doi: 10.1177/0040059917703962
- Gallucci, M. (2019). *GAMLj: General Analyses for Linear Models [Jamovi module]*. Available online at: <https://gamlj.github.io/>
- Gary, S., Lenhard, W., and Lenhard, A. (2021). Modelling norm scores with the cNORM package in R. *Psych* 3, 501–521. doi: 10.3390/psych3030033
- Gasteiger-Klicpera, B., and Klicpera, C. (2005). "Lese-Rechtschreibschwierigkeiten bei sprachgestörten Kindern der 2.-4. Klassenstufe [Reading and spelling difficulties in language-impaired children in grades 2 - 4]," in *Sprachentwicklungsstörungen früh erkennen und behandeln*, eds P. Arnoldy and B. Traub (Karlsruhe: Loeper), 77–95.
- Gebhardt, M., Heine, J.-H., Zeuch, N., and Förster, N. (2015). Lernverlaufsdiagnostik im Mathematikunterricht der zweiten Klasse: Raschanalysen und Empfehlungen zur Adaptation eines Testverfahrens für den Einsatz in inklusiven Klassen [Learning development diagnostics in second grade mathematics teaching: rapid analyses and recommendations for adapting a test procedure for use in inclusive classes]. *Empirische Sonderpädagogik* 7, 206–222. doi: 10.25656/01:11383
- Gebhardt, M., Jungjohann, J., and Schurig, M. (2021). *Lernverlaufsdiagnostik im förderorientierten Unterricht: Testkonstruktionen Instrumente Praxis [Learning process diagnostics in support-oriented teaching: Test constructions Instruments Practice]*. München: Reinhardt.
- Good, R. H., and Jefferson, G. (1998). "Contemporary perspectives on Curriculum-Based Measurement validity," in *Advanced Applications of Curriculum-Based Measurement*, ed M. R. Shinn (New York, NY: Guilford Press), 61–88.
- Hasbrouck, J., and Tindal, G. A. (2006). Oral reading fluency norms: a valuable assessment tool for reading teachers. *Read. Teach.* 59, 636–644. doi: 10.1598/RT.59.7.3
- Herrmann, S., Meissner, C., Nussbaumer, M., and Ditton, H. (2021). Matthew or compensatory effects? Factors that influence the math literacy of primary-school children in Germany. *Br. J. Educ. Psychol.* 92:e12462. doi: 10.1111/bjep.12462
- Hosp, M. K., Hosp, J. L., and Howell, K. W. (2016). *The ABC's of CBM: A Practical Guide to Curriculum-Based Measurement*. 2nd Edn. New York, NY: The Guilford Press.
- Hußmann, A., and Schurig, M. (2019). Unter der Norm - Kompetenz und Diagnostik in IGLU 2016 [Below the Norm - Competence and Diagnostics in PIRLS 2016]. *Empirische Sonderpädagogik* 11, 279–293. doi: 10.25656/01:18335
- Jaeuthe, J., Lambrecht, J., Bosse, S., Bogda, K., and Spörer, N. (2020). Entwicklung der Rechtschreibkompetenz im zweiten und dritten Schuljahr: Eine latente Transitionsanalyse zur Überprüfung theoretischer Annahmen [Development of spelling competence in the second and third school year: a latent transition analysis to test theoretical assumptions]. *Z. Erziehungswiss* 23, 823–846. doi: 10.1007/s11618-020-00959-5
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer New York. doi: 10.1007/978-1-4614-7138-7
- Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. 2nd Edn. New York, NY: Oxford University Press.
- Keuning, T., van Geel, M., and Visscher, A. (2017). Why a data-based decision-making intervention works in some schools and not in others. *Learn. Disabil. Res. Pract.* 32, 32–45. doi: 10.1111/ldrp.12124
- Kim, Y.-S. G., Petscher, Y., and Park, Y. (2016). Examining word factors and child factors for acquisition of conditional sound-spelling consistencies: A longitudinal study. *Sci. Stud. Read.* 20, 265–282. doi: 10.1080/10888438.2016.1162794
- Klauer, K. J. (2014). "Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik [Formative performance diagnostics: Historical background and further development to learning progress diagnostics]," in *Lernverlaufsdiagnostik*, eds M. Hasselhorn, W. Schneider, and U. Trautwein (Göttingen; Bern; Wien; Paris: Hogrefe), 1–17.
- KMK (2005). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4) [Educational standards in German for the primary level (grade 4)]*. Bonn: Kultusministerkonferenz.
- Kratochwill, T. R., and Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: randomization to the rescue. *Psychol. Methods* 15, 124–144. doi: 10.1037/a0017736
- Lenhard, A., Lenhard, W., and Gary, S. (2019). Continuous norming of psychometric tests: a simulation study of parametric and semi-parametric approaches. *PLoS ONE* 14, e0222279. doi: 10.1371/journal.pone.0222279
- Lenhard, A., Lenhard, W., Suggate, S., and Segerer, R. (2018). A continuous solution to the norming problem. *Assessment* 25, 112–125. doi: 10.1177/1073191116656437
- Lenhard, W., and Lenhard, A. (2014). (2014–2022). *Berechnung des Lesbarkeitsindex LIX nach Björnson. [Index of Readability by Björnson]*. Available online at: <http://www.psychometrica.de/lix.html>. Dettelbach: Psychometrica.
- Lenhard, W., Lenhard, A., and Schneider, W. (2017). *ELFE II - ein Leseverständnistest für Erst- bis Siebtklässler: Version II [ELFE II - a reading comprehension test for first to seventh graders]*. Göttingen: Hogrefe.
- Lervåg, A., and Hulme, C. (2010). Predicting the growth of early spelling skills: are there heterogeneous developmental trajectories? *Sci. Stud. Read.* 14, 485–513. doi: 10.1080/10888431003623488
- MacCallum, R. C., Zhang, S., Preacher, K. J., and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychol. Methods* 7, 19–40. doi: 10.1037/1082-989X.7.1.19
- Magis, D., Beland, S., Tuerlinckx, F., and De Boeck, P. (2010). *diffR: A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning*. Available online at: <https://CRAN.R-project.org/package=diffR>
- May, P. (1990). "Kinder lernen rechtschreiben: Gemeinsamkeiten und Unterschiede guter und schwacher Lerner [Children learn to write: Similarities and differences between strong and weak learners]," in *Das Gehirn, Sein Alphabet und andere Geschichten*, eds H. Brügelmann and H. Balhorn (Konstanz: Faude), 245–253.
- May, P., Malitzky, V., and Vieluf, U. (2019). *HSP+*. Stuttgart: Klett.
- McMaster, K., and Espin, C. (2007). Technical features of curriculum-based measurement in writing. *J. Spec. Educ.* 41, 68–84. doi: 10.1177/00224669070410020301
- Meijer, C. J. W., and Watkins, A. (2019). Financing special needs and inclusive education – from Salamanca to the present. *Int. J. Inclusive Educ.* 23, 705–721. doi: 10.1080/13603116.2019.1623330
- Mesquita, A., Carvalhais, L., Limpo, T., and Castro, S. L. (2020). Portuguese spelling in primary grades: complexity, length and lexicality effects. *Read. Writ.* 33, 1325–1349. doi: 10.1007/s11145-019-10012-5
- Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *J. Educ. Measurement* 29, 133–161. doi: 10.1111/j.1745-3984.1992.tb00371.x
- National Center on Response to Intervention (2010). "Essential Components of RTI – A Closer Look at Response to Intervention". Available online at: <https://files.eric.ed.gov/fulltext/ED526858.pdf> (accessed May 13, 2022).
- Naumann, C. L. (1987). *Rechtschreibwörter und Rechtschreibregeln: Hilfe für die Erarbeitung eines lerngruppenbezogenen Grundwortschatzes [Spelling words and spelling rules: Help for the development of a basic vocabulary related to learning groups] (2nd Ed.)*. Soest: Soester Verlagskontor.

- O'Connor, R. E., and Fuchs, L. S. (2013). "Responsiveness to intervention in the elementary grades: Implications for early childhood education," in *Handbook of Response to Intervention in Early Childhood*, eds V. Buysse and E. S. Peisner-Feinberg (Paul H Brookes Publishing Co.), 41–55.
- Peng, P., Fuchs, D., Fuchs, L. S., Elleman, A. M., Kearns, D. M., Gilbert, J. K., et al. (2019). A longitudinal analysis of the trajectories and predictors of word reading and reading comprehension development among at-risk readers. *J. Learn. Disabil.* 52, 195–208. doi: 10.1177/0022219418809080
- Prince, A. M. T., Yell, M. L., and Katsiyannis, A. (2018). Andrew F. v. douglas county school district (2017): the U.S. supreme court and special education. *Intervent. School Clinic* 53, 321–324. doi: 10.1177/1053451217736867
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. (Version 4.1) [Computer software]. Available online at: <https://cran.r-project.org>
- Reber, K., and Kirch, M. (2013). Richtig schreiben lernen. Kompetenzorientierter, inklusiver Rechtschreibunterricht [Learning to write correctly. Competence-oriented, inclusive spelling lessons]. *Praxis Sprache* 4, 254–257.
- Riehme, J. (1987). *Rechtschreibunterricht. Probleme und Methoden* [Teaching spelling. Problems and methods] (5. Aufl.). Frankfurt: Moritz Diesterweg.
- Rigby, R. A., and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Appl. Statistics* 54, 507–554. doi: 10.1111/j.1467-9876.2005.00510.x
- Robitzsch, A., Kiefer, T., and Wu, M. (2021). *TAM: Test Analysis Modules*. [R package]. Available online at: <https://CRAN.R-project.org/package=TAM>
- Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* [Textbook Test Theory - Test Construction]. 2nd Ed. Bern: Huber.
- Salaschek, M., Zeuch, N., and Souvignier, E. (2014). Mathematics growth trajectories in first grade: cumulative vs. compensatory patterns and the role of number sense. *Learn. Individual Diff.* 35, 103–112. doi: 10.1016/j.lindif.2014.06.009
- Schurig, M., Jungjohann, J., and Gebhardt, M. (2021). Minimization of a short computer-based test in reading. *Front. Educ.* 6:684595. doi: 10.3389/educ.2021.684595
- Sennlaub, G. (1985). *So wird's gemacht. Grundwortschatz – Auswahl und Arbeit* [How to do it. Basic vocabulary - Selection and work] (2. Aufl.). Heinsberg: Agentur Dieck.
- Seol, H. (2022). *snowIRT: Item Response Theory for jamovi*. [jamovi module]. Available online at: <https://github.com/hyunsooseol/snowIRT>
- Shinn, M. R. (1998). *Advanced Applications of Curriculum-Based Measurement*. New York, NY: Guilford Press.
- Souvignier, E. (2020). "Interventionsforschung im Kontext Schule [Research on Interventions in Schools]" in *Handbuch Schulforschung*, eds T. Hascher, T.-S. Idel, and W. Helsper (Wiesbaden: Springer), 1–17. doi: 10.1007/978-3-658-24734-8_9-1
- Souvignier, E., Förster, N., and Schulte, E. (2014). "Wirksamkeit formativen assessments - evaluation des ansatzes der lernverlaufsdiagnostik [Effectiveness of formative assessment - evaluation of the learning trajectory diagnostic approach]," in *Lernverlaufsdiagnostik*, eds M. Hasselhorn, W. Schneider, and U. Trautwein (Göttingen: Hogrefe), 221–237.
- Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S., and Haag, N. (eds.) (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* [IQB Education Trend 2016. Competences in german and mathematics at the end of grade 4 in the second country comparison.]. Waxmann.
- Strathmann, A., Klauer, K. J., and Greisbach, M. (2010). Lernverlaufsdiagnostik - Dargestellt am Beispiel der Entwicklung der Rechtschreibkompetenz in der Grundschule [Diagnosis of learning progression - illustrated by the example of the development of spelling competence in primary school]. *Empirische Sonderpädagogik* 2, 64–77. doi: 10.25656/01:9338
- Strathmann, A. M., and Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittmessung [Learning progress diagnostics: an approach to longer-term learning progress measurement]. *Zeitschrift für Entwicklungspsychol. Pädagog. Psychol.* 42, 111–122. doi: 10.1026/0049-8637/a000011
- Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journal. Q.* 30, 415–433. doi: 10.1177/107769905303000401
- The jamovi project (2022). *Jamovi*. (Version 2.3) [Computer Software]. Available online at: <https://www.jamovi.org>
- Thomé, G. (2003). "Entwicklung der basalen Rechtschreibkenntnisse, "Didaktik der deutschen Sprache, eds U. Bredel, H. Günther, P. Klotz, J. Ossner and G. Siebert-Ott (Paderborn: Ferdinand Schöningh), 369–379.
- Tindal, G. (2013). Curriculum-based measurement: a brief history of nearly everything from the 1970s to the Present. *ISRN Educ.* 2013, 1–29. doi: 10.1155/2013/958530
- van Geel, M., Keuning, T., Visscher, A. J., and Fox, J.-P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *Am. Educ. Res. J.* 53, 360–394. doi: 10.3102/0002831216637346
- van Ophuysen, S. (2010). "Professionelle pädagogisch-diagnostische Kompetenz - eine theoretische und empirische Annäherung [Professional pedagogical-diagnostic competence - a theoretical and empirical approach]," in *Jahrbuch der Schulentwicklung: Band 16: Daten, Beispiele und Perspektiven*, eds N. Berkemeyer, W. Bos, H. G. Holtappels, and N. McElvany (Weinheim: Juventa), 203–234.
- Vaughn, S., Linan-Thompson, S., and Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Except. Children* 69, 391–409. doi: 10.1177/001440290306900401
- Verhoeven, L. (2000). Components in early second language reading and spelling. *Sci. Stud. Read.* 4, 313–330. doi: 10.1207/S1532799XSSR0404_4
- Voncken, L., Albers, C. J., and Timmerman, M. E. (2019). Model selection in continuous test norming with GAMLSS. *Assessment* 26, 1329–1346. doi: 10.1177/1073191117715113
- Voß, S., and Blumenthal, Y. (2020). Assessing the word recognition skills of german elementary students in silent reading—psychometric properties of an item pool to generate curriculum-based measurements. *Educ. Sci.* 10:35. doi: 10.3390/educsci10020035
- Voß, S., Blumenthal, Y., Mahlau, K., Marten, K., Diehl, K., Sikora, S., et al. (2016). *Der Response-to-Intervention-Ansatz in der Praxis: Evaluationsergebnisse zum Rügener Inklusionsmodell* [The Response-to-Intervention Approach in Practice: Evaluation Results on the Rügen Inclusion Model]. Münster, New York, NY: Waxmann.
- Voß, S., Sikora, S. and Mahlau, K. (2017). Vorschlag zur Konzeption eines curriculumbasierten Messverfahrens zur Erfassung der Rechtschreibleistungen im Grundschulbereich [Proposal for the conception of a curriculum-based Measurement procedure for the assessment of spelling performance at primary school level]. *Empirische Sonderpädagogik* 9, 184–194. doi: 10.25656/01:15029
- What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook (Version 4.1)*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Available online at: <https://ies.ed.gov/ncee/wwc/handbooks>
- Wilbert, J., and Linnemann, M. (2011). Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik [Criteria for analyzing a test measuring learning progress]. *Empirische Sonderpädagogik* 225–245. doi: 10.25656/01:9325
- Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. MESA Press.
- Wright, K. D., and Oshima, T. C. (2015). An effect size measure for raju's differential functioning for items and tests. *Educ. Psychol. Measurement* 75, 338–358. doi: 10.1177/0013164414532944
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl. Psychol. Measure.* 8, 125–145. doi: 10.1177/014662168400800201
- Zachary, R. A., and Gorsuch, R. L. (1985). Continuous norming: implications for the WAIS-R. *J. Clin. Psychol.* 41, 86–94. doi: 10.1002/1097-4679(198501)41:1<86::AID-JCLP2270410115>3.0.CO;2-W



OPEN ACCESS

EDITED BY

Erica Lembke,
University of Missouri,
United States

REVIEWED BY

Jamie Capal,
University of North Carolina at Chapel Hill,
United States
Niki Pandria,
Aristotle University of Thessaloniki, Greece

*CORRESPONDENCE

Gianluca Merlo
✉ gianluca.merlo@itd.cnr.it

SPECIALTY SECTION

This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

RECEIVED 13 May 2022

ACCEPTED 28 December 2022

PUBLISHED 18 January 2023

CITATION

Merlo G, Chifari A, Chiazese G, Denaro P,
Firrera N, Savio NL, Patti S, Palmegiano L,
Taibi D and Seta L (2023) The BEHAVE
application as a tool to monitor inclusive
interventions for subjects with
neurodevelopmental disorders.
Front. Psychol. 13:943370.
doi: 10.3389/fpsyg.2022.943370

COPYRIGHT

© 2023 Merlo, Chifari, Chiazese, Denaro,
Firrera, Savio, Patti, Palmegiano, Taibi and
Seta. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

The BEHAVE application as a tool to monitor inclusive interventions for subjects with neurodevelopmental disorders

Gianluca Merlo^{1*}, Antonella Chifari¹, Giuseppe Chiazese¹,
Paola Denaro¹, Noemi Firrera², Nicola Lo Savio², Simona Patti²,
Luisa Palmegiano², Davide Taibi¹ and Luciano Seta¹

¹Istituto per le Tecnologie Didattiche, Consiglio Nazionale delle Ricerche, Palermo, Italy, ²Istituto Tolman, Palermo, Italy

In the last few years, many educational and therapeutic interventions for young people with neurodevelopmental disorders are based on systematic monitoring of the outcomes. These interventions are typically conducted using single-case experimental designs, (SCEDs) a set of methods aimed at testing the effect of an intervention on a single subject or a small number of subjects. In SCEDs, an effective process of decision-making needs accurate, precise, and reliable data but also that caregivers and health professionals can gather information with minimal effort. The use of Information Communication Technologies in SCEDs can support the process of data collection and analysis, facilitating the collection of accurate and reliable data, providing reports accessible also by non-experts, and promoting interactions and sharing among clinicians, educators, and caregivers. The present paper introduces the BEHAVE application, a web-based highly customizable application, designed to implement SCEDs, supporting both data collection and automatic analysis of the datasets. Moreover, the paper will describe two case studies of kindergarten children with neurodevelopmental disorders, highlighting how the BEHAVE application supported the entire process, from data collection in multiple contexts to decision-making based on the analysis provided by the system. In particular, the paper describes the case studies of Carlo and Dario, two children with severe language and communication impairments, and the inclusive education interventions carried out to maximize their participation in a typical home and school setting increasing their mand repertoire. Results revealed an increase in the mand repertoire in both children who become able to generalize the outcomes to multiple life contexts. The active participation of the caregivers played a crucial role in the ability of children to use the learned skills in settings different from the ones they were learned in.

KEYWORDS

neurodevelopmental disorders, applied behavior analysis, communication impairments, technologies for monitoring, inclusive intervention

1. Introduction

Single-case experimental design (SCED) is an expression used to indicate a class of research designs, characterized by repeated observations of a single entity (generally an individual, but can be also a group, a classroom, a school, or a hospital) in a fixed period during at least a variable is manipulated, generally the treatment.

SCEDs are often contrasted with randomized control trials (RCTs), in which two or more groups, control vs. experimental groups, are compared with the aim to establish the effects of an intervention using standardized and validated instruments. For many years, RCTs were the “gold standard” for experimental studies. Recently, a growing number of works have criticized the general usefulness of these clinical trials to discern the impact on the individual health of some specific treatments (Jacobson and Christensen, 1996; Westen and Bradley, 2005; Wachtel, 2010; Branch, 2014; Perone, 2019). An RCT study can capture the “average” effect but is not suitable to determine the causal/functional relationship between a treatment and observed change for one individual: “managers and trialists may be happy for treatments to work on average; patient’s doctors expect to do better than that” (Evans, 1995; cited by Vlaeyen et al., 2020, p. 659). SCEDs are especially suitable for establishing evidence of intervention efficacy and conducting pilot investigations, also for large-scale causal studies (Smith, 2012; Natesan Batley et al., 2020). Although the methodology underpinning SCED has a long history, at least since the work of experimental psychology based on the behavioral approach and operant conditioning (Skinner, 1938; Sidman, 1952, 1960; Skinner, 1956; Shapiro, 1966) recent epistemological and methodological developments have brought this type of study back into the mainstream.

The use of SCEDs is historically linked to the study of human and animal behavior and the search for causal or functional relationships between the manipulation of the subject’s living environment and observable changes in his/her behaviors (Tate and Perdices, 2019; Vlaeyen et al., 2020; Kazdin, 2021). This methodology is therefore frequently used when it comes to testing the effectiveness of therapeutic interventions in developmental disorders. For example, in a recent systematic review of behavior analytic interventions for young children with intellectual disabilities (Ho et al., 2021) of the 49 studies included, only three (6%) were group-design studies, and the rest used single-case design methodology. Moreover, SCED appears as the prevalent methodology (Cannella-Malone et al., 2021) to monitor problem behaviors related to the most common neurodevelopmental disorders (Horner et al., 2005; Odom et al., 2005; Cook and Cook, 2013; Cook and Odom, 2013).

For this type of issue, the interventions are often tailored to the specific characteristics of the individual and based on the assumption that behavioral change can be the result of a learning process. The comparison between groups can be affected by important biases, related to the difficulty to have homogeneous groups, the intervention of spurious variables, the

interpretation of the results in terms of efficacy on a single individual, and the translation of correlation in a functional relationship.

SCEDs can also be affected by the risk of bias. The most frequent risks are related to the inability to conceal certain elements of the research design from study participants, researchers, and individuals collecting outcome data. Another frequent risk of bias is related to the lack of clear documentation of fidelity to the experimental procedures (Smith et al., 2022).

The use of ICTs in SCEDs is relevant to reduce the risk of bias supporting the process of accurate and reliable data collection (Spachos et al., 2014). Moreover, ICTs could promote the generation of reports accessible also by non-experts, and the interaction and sharing among the different caregivers.

The present paper introduces the BEHAVE application as a tool to promote the culture of evidence-based principles both in clinical and educational contexts, facilitating the process of monitoring and management of the problem behavior linked to neurodevelopmental disorders (NDDs) and providing users with an easy way to gather data and evaluate the effect size of the behavioral interventions.

In particular, the paper will introduce the BEHAVE web application as a technological tool supporting an ABA inclusive intervention applied to two kindergarten children with autism and severe language and communication impairments. First, theoretical points of departure about autism spectrum disorder (ASD) and language and communication impairments will be described. Then the paper will describe the two mentioned case studies, highlighting how the BEHAVE application supported the entire process, from data collection in multiple contexts to decision-making based on the analysis provided by the system.

2. Autism spectrum disorder and language and communication impairments

According to DSM-5 (American Psychiatric Association, 2013), the diagnostic class of neurodevelopmental disorders (NDDs) comprises disorders that arise during the developmental period characterized by personal, social, scholastic, or occupational difficulties. The phenotypes of NDDs are very heterogeneous including for example intellectual disabilities, autism spectrum disorders (ASD), attention-deficit/hyperactivity disorders, communication disorders, neurodevelopmental motor disorders, and specific learning disorders. The etiology of these disorders is considered multifactorial (Guo et al., 2018) involving, among others, genetic, perinatal, endocrine, and psychosocial risk factors. The prevalence of NDDs is highly variable changing as a function of the disorder typology, socioeconomic factors, and sex. For example, intellectual disabilities range from 0.3 to 8% according to the severity (Simonoff, 2015), speech disorders from 2 to 31% (Norbury and Paul, 2015), and autism spectrum disorders from 0.6 to 1% in developed countries (Williams et al.,

2006; Matson and Kozlowski, 2011) to 3% in South Korea and Japan (Elsabbagh et al., 2012).

One of the most studied NDDs is ASD. According to a behavioral perspective, ASD is a syndrome characterized by behavioral deficits and excesses, which have a neurological basis, but can be modified as a result of specific interactions with the environment (Martin and Pear, 2019). ASDs are biologically determined neurodevelopmental conditions that generally begin in the first 3 years of life and accompany the individual throughout the life cycle.

The predominantly affected areas are those related to communication and social interaction and the presence of restricted and repetitive patterns of behavior, interests, or activities (American Psychiatric Association, 2013). Deficits in these areas can be expressed in very different ways from one person to another, can vary over time depending on the interaction with the context, and can foster the emergence of different types of emotional, social, and behavioral disorders. Deficiencies in communication skills are some of the most common deficits in people with autism spectrum disorders (Peeters and Gillberg, 1999). Communication skills are fundamental for good children's social interaction within their living environment. Deficiencies in these skills can lead to the emergence of problem behaviors, which sometimes represent substitute modes of communication (Cooper et al., 2007; Moderato and Copelli, 2010; Sundberg and Sundberg, 2011).

Several studies have stressed the importance of early identification of subclinical signs of autism to facilitate access to early treatment and improve prognosis (Daniels et al., 2014; Costanzo et al., 2015). Indeed, untimely interventions put children at risk of developing sleeping disorders (Kamara and Beauchaine, 2020), being physically and sexually abused (Balogh et al., 2001; Reiter et al., 2007; Jawaid et al., 2012), committing suicide (Lunsky, 2004) or violent crimes (Lundström et al., 2014).

Behavioral interventions based on applied behaviour analysis (ABA) now represent some of the best treatments available for developing communication skills in people with ASD (Peters-Scheffer et al., 2011; Reichow et al., 2018; Makrygianni et al., 2018; Yu et al., 2020).

ABA is a science based on learning principles that aim to predict and influence behavior to promote socially meaningful behavior (Cooper et al., 2007). ABA enables the implementation of individualized interventions to develop deficient skills and reduce barriers to learning in individuals with autism spectrum disorders through the identification and modification of contextual variables that influence behavior and the application of systematically applied procedures.

Starting from the seminal work of Skinner (1957), behavior analysis has developed a functionalist approach to language analysis and verbal relations (Presti et al., 2002), which has allowed the definition of procedures that are effective in promoting communication skills in people with cognitive delays and disabilities (Presti et al., 2002; Carbone et al., 2010). Skinner (1957) considers language as learned behavior and defines verbal

communication as an operant behavior of a speaker reinforced through the mediation of a listener, who has learned in the verbal community to provide appropriate consequences to the speaker's behavior. From this perspective, the main focus is on the function as well as the form of the verbal relationship.

A taxonomy of responses called verbal operants is defined starting from the identified behavioral function. This system of analysis and classification has important implications (Sundberg and Michael, 2001) for the development of language and communication in individuals with autism spectrum disorder, both through the teaching of vocal language and through the use of alternative augmentative communication (AAC) systems, such as the use of gestures and equipment for the partial or total, temporary or permanent, compensation of severe difficulties in the emission of vocal language (Cafiero, 2009). The acquisition of verbal behavior promotes the development of cognitive, scholastic, and social skills (Sundberg and Michael, 2001).

Among verbal operants, mands are fundamental for the development of language and social interactions. Skinner (1957) defines the mand as "a verbal operant in which the response is reinforced by a characteristic consequence and is therefore under the functional control of relevant conditions of deprivation or aversive stimulation" (p. 35–36). Mand is a type of verbal behavior controlled by motivational operations. It allows the speaker to request what he needs and wants and is usually among the first forms of children's communication (Sundberg, 2008). Hand signs (Carbone et al., 2010), pictures exchange communication system (Jurgens et al., 2009), and voice imitation training (Ross and Greer, 2003) are successful examples of strategies for vocal imitation mand teaching.

3. Technology to assess, monitor, and treat NDDs

Technologies-based monitoring practices are conceived as an evidence-based data collection process able to capture dynamic changes in psychological, cognitive, and behavioral outcomes and to support customized individual interventions (Bentley et al., 2018) in clinical practices. The spread of smart devices (such as mobile and wearable devices) has opened up scenarios where it is possible to reach the subjects by assessing them through observations in their natural environment. In particular, the development of new digital health solutions and services has allowed therapists to assess, monitor, and treat the patient by offering a more suitable and feasible digital intelligent health service.

A recent systematic review by Valentine et al. (2020) identifies how smart devices like tablets, smartphones, and wearable devices may be combined with apps, gaming applications, and video modeling behavioral training activities to support the assessment, diagnosis, treatment, and monitoring processes. Examples of clinical services supported by technologies are virtual reality assessment/therapy, telehealth assistance, computer-based

assessment/therapies, and monitoring across multiple NDDs, especially ASD and ADHD disorders. The review also highlights the positive technological impact on clinical effectiveness, economic cost–benefit, and the user feasibility and acceptability of technology (Valentine et al., 2020).

The introduction of behavioral tracking with smart devices (Shiffman et al., 2008; Shingleton et al., 2016; Wichers and Groot, 2016), and the use of wearable biosensors for physiological assessment (Schlier et al., 2019) underpins the collection of repeated systematic observation over time (Bentley et al., 2018) and the advancement in the quantitative techniques used for SCED has enhanced and facilitated the interpretation of outcomes, the communication of the results between different subjects, and comparing the results using common, quantitative approaches (Barnard-Brak et al., 2022).

In the last few years, one of the prevalent technologies for developing behavioral monitoring applications according to SCED has been touch screen-based smart devices, and this trend is likely to continue in the coming years (Saini and Roane, 2018).

The development of these applications has followed two different paths. First, there are numerous research projects that have created tools (often with open-source code releases) designed with the purpose of providing practitioners with software to support data-driven decisions. Some non-exhaustive examples are the WHAAM application for monitoring subjects with ADHD (Spachos et al., 2014), BDataPro that supports behavioral data collection and visual analysis of the same (Zheng et al., 2022), BASE for implementing behavioral interventions at school (Chiazzeze et al., 2019), and the AHA application for monitoring and evaluating the attention skills of subjects with ADHD engaged in the use of an augmented reality-based solution to stimulate reading and writing skills (Tosto et al., 2021).

Second, many software houses developed paid applications that could support a specific niche of health professionals working with neurodevelopmental disorders and, more generally, with behavioral disorders. An overview of existing commercial systems and their peculiarities is provided by Merlo (Merlo et al., 2020).

What emerges from the analysis of these applications is that, in general, they are inflexible and not customizable since they are often anchored to a specific theoretical and methodological approach. Moreover, in most cases, the software do not provide sophisticated tools for automatic data analysis, leaving users with the task of visually analyzing the data or exporting the collected dataset for later autonomous analysis.

The BEHAVE system that will be presented below aims to address these limitations by providing its users with a highly customizable tool so that each practitioner can use it in accordance with his or her own approach. In addition, the system is capable of generating reports based on automatic statistical analyses that can be understood even by those without strong statistical skills.

The BEHAVE application is a tool that was created as the main output of a project funded by the European Commission in 2017 (2017-1-IT02-KA201-036540) inside a KA2 Strategic Partnership for school education Erasmus+. More in-depth, the project was

aimed at enhancing the experience and expertise of health professionals, teachers, parents, and caretakers in the management of behavioral interventions, according to the idea that the SCED methodology is a good practice to measure processes and procedures and to support the implementation of evidence-based practices, clarifying what are the most effective strategies case by case.

As mentioned above, the most innovative feature of the BEHAVE application is the opportunity to support the management of selected problem behaviors through the creation of custom measures, the operational definition of behavior, the collection of behavioral data, and the comparison between phases (e.g., baseline and intervention) through both the visual comparison of data represented by scatter plots and statistical analyses that are generated automatically.

In particular, the application can identify the best algorithm of effect size among those developed by Parker et al. (2011) and Allison and Gorman (1993), returning the effect size of the intervention displayed simply and clearly, through a speedometer that guides to the meaning of the data collected even the least experienced user. Providing users with the data and their analysis in a simple and accessible way could thus facilitate the introduction of evidence-based scientific approaches in multiple contexts beyond the clinical one.

In the BEHAVE application, users can combine six different types of questions, called items, to collect data about behaviors. One of these is the “Direct observation item” useful to collect data on how long or how often a certain behavior occurs during the observation. For example, the frequency, duration, and intensity of specific behavior can be assessed when a student interrupts a class, leaves his seat, raises his hand, yells out an answer, or asks to go to the bathroom. Direct observation items support the frequency, duration, and interval recording. The choice of the recording procedure depends on the typology of behavior that caregivers want to observe.

Other measures are:

- The “Choice item” is useful when the question includes one or more answers among a group of predefined answers. A famous example of choice item with one answer allowed is the Likert scale, a scale composed of items to whom respondents must specify their level of agreement or disagreement on a symmetric agree-disagree scale.
- The “Number item” are questions that can be answered only with numbers. For example, these items can be useful to count how many times a child threw a pencil or other object at other pairs. A simple numeric item such as “How many times the child threw the pencil?” can be created to obtain a numerical answer.
- “Range items” are similar to number items but instead to accept any integer value, they accept only values included between a minimum and a maximum.
- “Four quadrant item” is used to measure at the same time two different dimensions. The measure is composed of two

dimensions displayed on a cartesian plane. Users have to choose how to position themselves in the two dimensions.

- “Text items” are aimed at gathering qualitative data about a phenomenon through the generation of open-ended questions.

The BEHAVE application allows users to combine different already existing measures to create their favorite combination of items. Users can import and export the created measures and share them with others to support a community of practice.

Finally, the BEHAVE application makes available a set of education support tools to coach the users through video tutorials within a Moodle course, a user guide to introduce the BEHAVE functionalities, and contextual help during the navigation of the application. The BEHAVE application is free of charge and it is accessible at the URL: <https://www.behaveproject.eu/>.

4. Case studies

Carlo is a 3-year-old boy diagnosed with autism spectrum disorder who does not have vocal language. Dario is a 3 years and 5 months boy with a diagnosis of agenesis of the corpus callosum and autism spectrum disorder who presents unintelligible vocal language due to phonetic-phonological difficulties. The diagnoses were made at the territorial child neuropsychiatry services. The two children attend nursery school and have three weekly ABA therapy sessions of 1 h each. The initial assessment of the functional skills of the two children was carried out by cognitive behavioral psychotherapists through the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP; Sundberg, 2008). The assessment highlighted important deficiencies in verbal skills, play skills, and social skills in both children. As a consequence, the education intervention plans focused primarily on the increase of both the frequency and variety of unprompted mands emitted by the children.

The study was carried out following the ethical principles and codes of the institution that delivered the treatments which are based on national and international ethics codes (e.g., the code of ethics of Italian psychologists and the BACB's Ethics Code for Behavior Analysts). Accordingly, approval by an ethics authority was not required. The informed consent form has been signed by the parents of the children involved in the study before the start of data collection.

4.1. Definition of the target behavior

The dependent variable measured in this study was the number of mands made by Dario and Carlo for accessing edible and dynamic stimuli, as well as tangible reinforcing objects available in the environment. In the case of Dario, the presence of vocal language facilitated the development of vocal mand even if the vocalization was not used to make requests and was often not

very intelligible. In the case of Carlo, alternative augmentative communication based on signs has been implemented to face the absence of vocal language. The definition of the target behavior is one of the first steps in the BEHAVE application. The therapist has to insert the behavior to be observed in the most accurate way specifying an operative description, the place, and the setting in which the behavior occurs. Figure 1 shows the definition of the target behaviors for the two case studies.

4.2. Measure creation and data recording

In the present study, data about the behaviors were collected through direct observations. In particular, a frequency direct observation measure was created with the BEHAVE application to measure how many times the number of mands occurred during the one-hour baseline and intervention sessions. The BEHAVE application was used to facilitate data collection in different contexts. In fact, in addition to the clinical context, the application facilitated the sharing of results between the figures involved in the therapeutic process, promoting the active participation of caregivers and maintaining high motivation for treatment. The BEHAVE application allowed the therapist to create a shareable URL to invite the children's parents to collect data in the home context.

4.3. Design and procedure

The dependent variable measured in this study was the number of correct mands issued on 50 given occasions (10 × 5 training items) within a 60-min therapy session by Dario and Carlo.

An AB design was defined within the BEHAVE application and employed to evaluate the effectiveness of the intervention by comparing data gathered during the treatment and baseline phases. Data were collected through direct observations with the BEHAVE application in a different context by the therapists and the parents. During the therapy sessions, 35 observations were made, divided into 5 baseline and 30 training observations. The generalization phase included 10 observations during home sessions with parents. The minimum empirical value assumed by the dependent variable was 0 and the maximum value 50.

The intervention progress was periodically monitored through the scatter plots and automated analyses provided by the BEHAVE application.

4.3.1. General procedure

According to Carbone et al. (2010), the first step of the teaching procedure is the selection of five items as target mand for each participant. This selection is the result of a previous assessment of the child's motivation in which edible, dynamic stimuli and tangible objects (toys) are proposed to the child. The teaching sessions included 50 trials delivered in an individual

Edit observation

Behaviour to be observed*	Frequency of correct mands (unprompted)
Description*	Number of correct mands (vocal/sign) issued on 50 given occasions (10 x 5 training items) within a 60-minute therapy session.
Place	Therapy
Setting	Istituto Tolman
Measure*	Frequency
Schedule observation dates*	<input type="checkbox"/> OFF

FIGURE 1
Screenshot of the target behaviors form of the BEHAVE application.

setting: 10 opportunities (trials) were offered to request each motivating item, for a total of 50 trials per session. Targets were presented in a randomized rotation. At the beginning of each trial, the therapist presented the desired item to the child making him aware of the availability of the reinforcer. If the child did not show motivation toward the presented item within 5 s, the therapist presented the next target of the rotation. Otherwise, the therapist started a specific teaching procedure to maximize the child's responses. [Figure 2](#) details the structure of a single therapeutic session with mand training so as to facilitate the replicability of the study.

4.3.2. Baseline

For the present study, the baseline data was gathered during 5 sessions in which each child had 50 opportunities to make mand (10 opportunities for each item) for their reinforcers within one therapy session. The mand emitted within 5 s of the presentation of the stimulus is considered a correct response. The response was considered incorrect in all other cases, including non-response. Both children at baseline scored zero.

4.3.3. Mand training: Vocal prompt/physical prompt and delay prompt

In the present study, the mand training provided both children with a procedure that included the evocation of the motivation to request the desired object, the use of prompts to model the correct request, and the delivery of the stimulus corresponding to the motivation as a specific reinforcer. For Dario, the vocal prompt was used to model vocal communication, while for Carlo the

physical prompt was used to model communication through signs. The training phase was initially carried out in the clinical setting by psychologists specialized in cognitive-behavioral psychotherapy and ABA. Moreover, the therapists implemented parent training to promote the generalization of learning in the natural environment.

When the child showed motivation for the items, the therapist provided the prompt (vocal for Dario and physical for Carlo's sign manuals). If the child emitted the mand (in echoic for Dario and in mimetic for Carlo), the therapist waited 5 s (5-s delay). If during the 5-s delay the child performed the mand independently, the therapist delivered the desired item immediately and left it available for 30 s. Otherwise, the therapist repeated the procedure from the beginning. If the child did not make the correct response for three sequences of presentation of the procedure, the therapist delivered the desired item anyway (with a lower magnitude).

4.3.4. The generalization of the mand skill with carers

When each child learned to emit the target mands for all of the 5 targets and reached 80% of correct mands in 3 consecutive therapy sessions (end of the intervention phase), parents were involved in 2 sessions of parent training and then they started to collect data through the BEHAVE application for 10 home sessions to monitor the generalization of the skill in the family's natural context.

During parent training, parents were trained in the use of the behavior application and in manipulating the motivation of children to create opportunities for mand to be emitted. To do

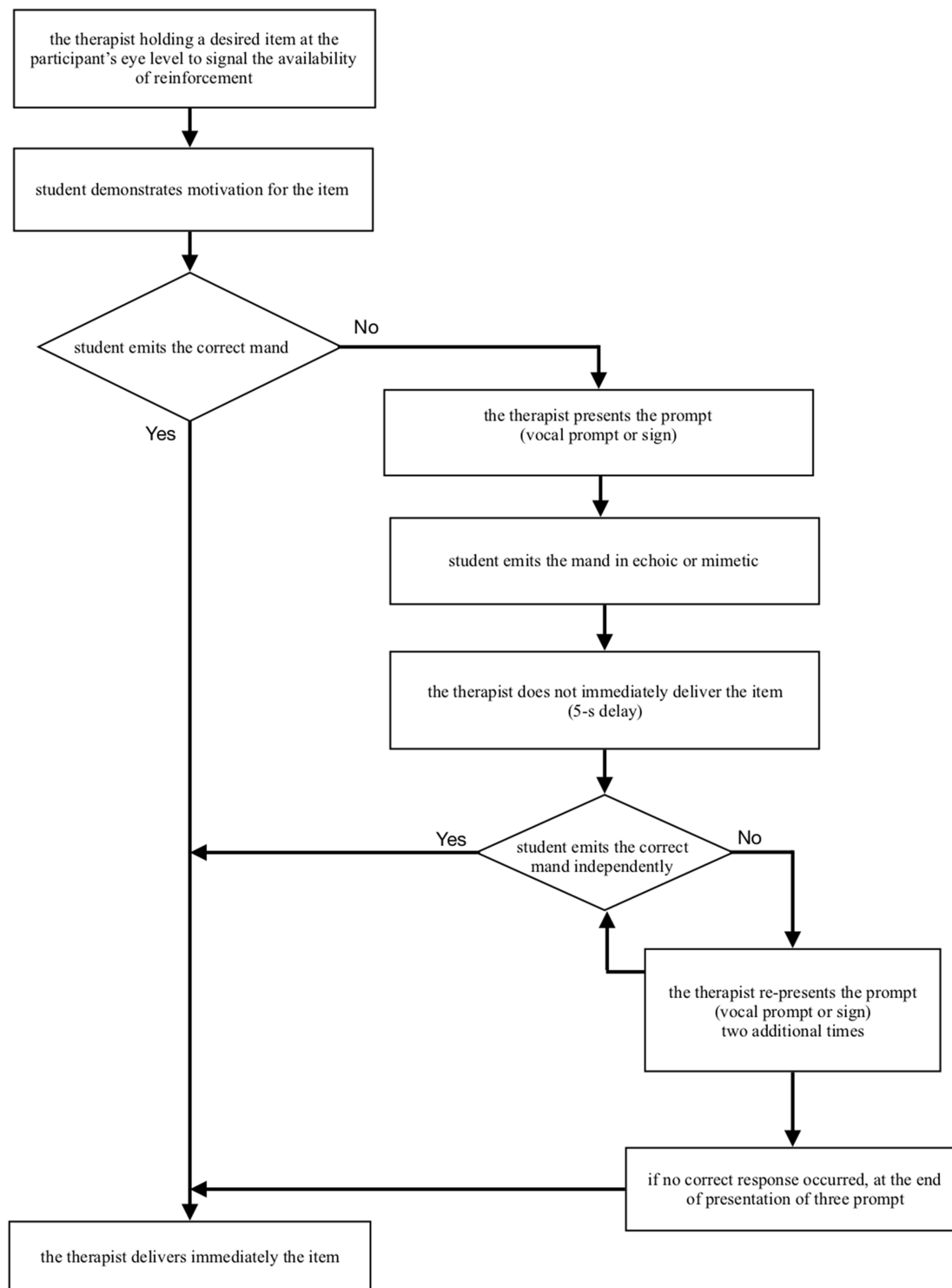


FIGURE 2
Flow chart describing the structure of a single therapy session with mand training.

this, the parents were asked to create one-hour play sessions, in which the trained items were made available (on sight but not directly accessible), and the correct requests emitted by the children were recorded through BEHAVE application.

During periodic monitoring of data with the BEHAVE application, therapists noticed that while Dario immediately generalized the mand skill learned in the therapy session, Carlo was not able to generalize the skill independently. Carlo did not

issue mand after three sessions of structured play at home and had therefore obtained a score of zero. For this reason, the therapists contacted Carlo's parents to support them in teaching the mand to their child at home.

After the generalization phase at home, to evaluate the ecological impact of the training, the parents were asked to complete a short interview that investigated the following areas: (1) emission at the home of spontaneous mands for trained targets also when the motivating objects are not on sight in the environment; (2) emission in other contexts of spontaneous mands for trained targets; (3) contexts in which spontaneous mands for trained targets emerge; (4) increase in the communicative repertoire of communicative intentionality (presence of requests or approximations of requests for new items).

In conclusion, both children learned the mand in training and generalized them also at home with the caregivers, although at different times and modalities (Figure 3).

5. Results

The results showed an increase in the mand repertoire in both children and a generalization of this ability in various contexts.

Dario emits the first correct mands in the 8th therapy session and reaches 80% of the correct mands in the 20th session. Stability in the results is achieved by session 31. Carlo emits 80% of correct mands at the 32nd therapy session. The 100% of correct mands is achieved by session 33.

In addition, monitoring the data collected by caregivers through the BEHAVE application, the therapists observed that the results obtained in the generalization phase were different for the two children. Dario has immediately generalized the skills learned during the intervention, showing continuous growth in the frequency of mand issued at home. Carlo, on the other hand, generalized the skills acquired in the family context after a specific parent training.

The following paragraphs describe in detail the results obtained comparing baseline with intervention, and intervention with generalization.

5.1. Baseline versus intervention

A quasi-experimental single-case AB design was performed, using the parametric method of Allison and Gorman (1993) as a method of analysis to evaluate the effect of the intervention. This method was suggested and applied automatically by the BEHAVE application according to the number of observations made in the baseline and treatment phases. The basic assumption for carrying out this analysis has been respected ($cov > 0$). The parametric analysis was applied to evaluate the effect of a possible change in frequency levels by considering the effect of the treatment under two different aspects: the first is the potential effect of treatment on the average change in frequency of mands between the baseline and treatment phases (effect on the levels); the second is the potential effect concerns the change in trajectory that the target behavior can assume

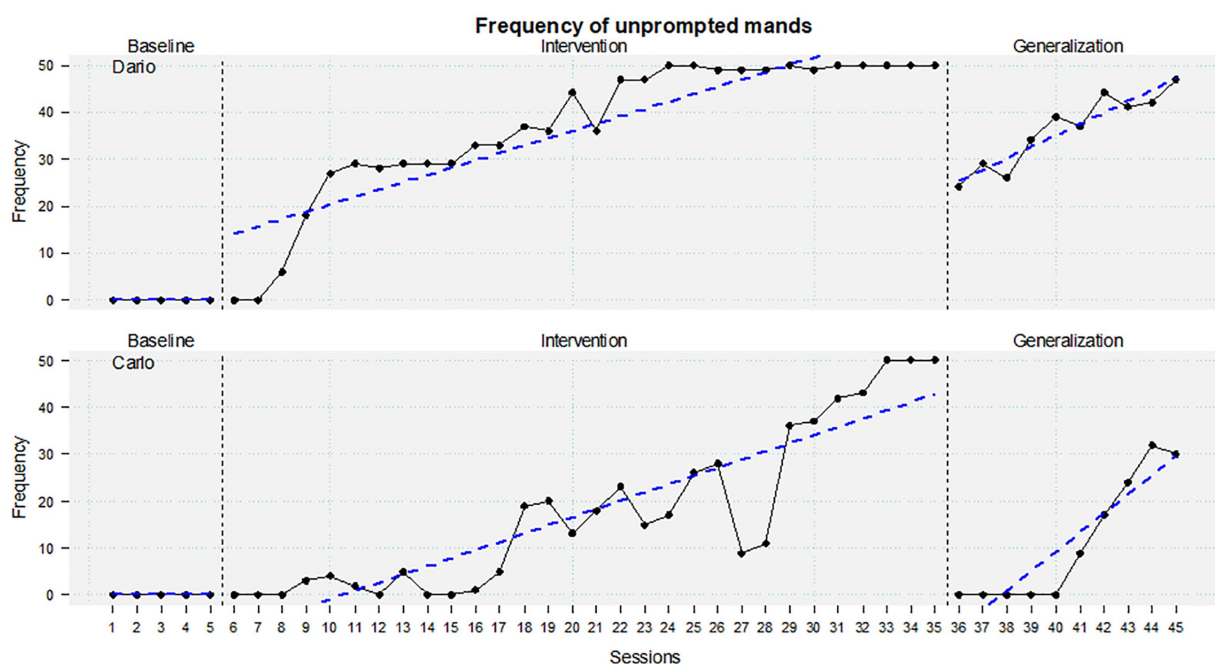


FIGURE 3

Scatterplot comparing baseline, intervention, and generalization of the unprompted mands for Carlo and Dario.

when passing from the baseline to the treatment phase (effect on the slopes). These two effects, indicating the frequency of the behavior changes in average terms and trajectory, have been estimated also considering the natural trend of the behavior of concern. The expression “natural trend” is intended to mean how the behavior of concern would have evolved naturally without any intervention. At this point, this estimate is subtracted from the observations made, obtaining a new variable called the de-trend score. This new variable expresses the frequency levels of the behavior, net of the variations that it would have naturally assumed over time. Using the de-trend score, it is possible to estimate the effect of the treatment, keeping the natural trend of our dependent variable under control.

During the treatments, both children show a progressive and linear increase in the number of correct mands over time.

The effect size values indicate a large impact of the treatment in increasing the mand repertoire during the therapy sessions both for Dario ($r=0.95$) and Carlo ($p=0$, $r=0.91$). The BEHAVE application summarizes the data relating to the values of the effect sizes through the speedometers shown in Figure 4. The treatment had a significant effect for both Dario ($R^2=0.94$, $F(2, 32)=143.33$, $p<0.05$) and Carlo ($R^2=0.83$, $F(2, 32)=77.99$, $p<0.05$).

5.2. Intervention versus generalization

The results obtained in the generalization phases were different for the two children. Dario has immediately started to generalize the skills learned during the intervention, showing a progressive and linear increase in the frequency of mand issued at home. Carlo, on the other hand, generalized the

skills acquired in the family context after specific parent training.

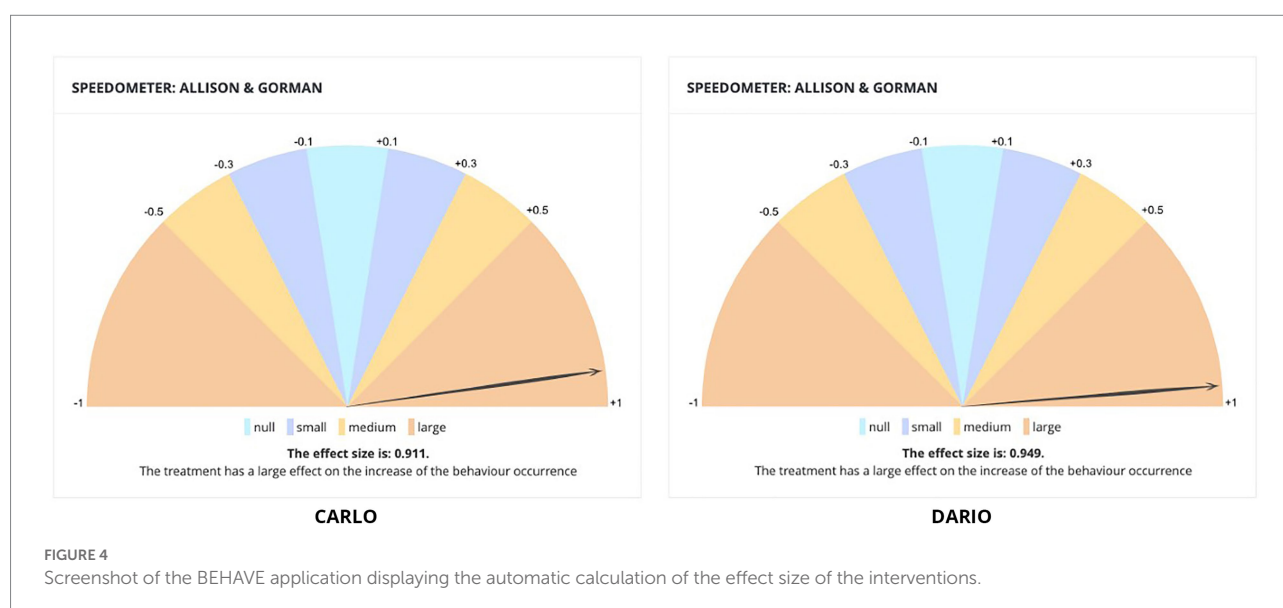
During the 10 generalization sessions at home, the mand ability showed a growing trend in both children.

The analysis of the parents' interviews showed that the children still use the repertoire of mands learned with their parents, grandparents, and in the school context. Dario show mands even when the motivating object is not in sight, while Carlo only when the items he wants are visible in his environment. Finally, the parents report an increase in communicative intentionality, reporting that children show greater interest in objects and people in the environment and begin to spontaneously manifest approximations of requests for motivating objects on sight not trained yet.

6. Discussion

Bringing up children with NDDs involves several issues that affect all the educational agents involved: parents and family of origin, caregivers, and teachers. Parents of children with disabilities generally have higher levels of stress than parents of typically developing children (Meadan et al., 2010; Karst and Van Hecke, 2012; Dykens, 2015). Similarly, working with children with NDDs can expose teachers, especially special needs teachers, to high levels of stress (Ghani et al., 2014) and unpleasant emotions which can eventually lead to burnout (Abel and Sewell, 1999; Richards et al., 2018). In general, teachers are stressed the most by behaviors that can hurt others, such as kicking, hitting, or biting (Amstad and Müller, 2020).

Many studies showed that ABA inclusive interventions may be effective in supporting the skills enhancement of children with



ASD (e.g., Camargo et al., 2014) reducing risk factors not only for the child but for the entire family and educational network that cares for the child.

The present study explored the possibility to use the BEHAVE web application as a technological tool supporting an ABA inclusive intervention on two children with severe language impairments. Results revealed that the interventions significantly increased the mand repertoire of both children and that the technological solution facilitated the collection of reliable data not only by experienced therapists but also by caregivers in the home context. As pointed out in the literature (Aktaş and Ciftci-Tekinarslan, 2018), this generalization seems to have been facilitated by the active participation of the carers, trained through careful parent training both to implement the procedure and to record the data with BEHAVE application, which made their collection easier. The BEHAVE application has facilitated the gathering of data in various contexts and the sharing of the results between the various figures involved.

Moreover, in the case of Carlo, the BEHAVE application allowed therapists to quickly identify the sign of difficulties in the generalization of results at home and to intervene promptly to guide his parents to maintain the skills learned in the clinical setting. These results are consistent with findings from other studies enhancing psychological interventions for subjects with NDDs through technologies (e.g., Hetzroni and Tannous, 2004). Artoni and colleagues, for example, identified positive improvements in all children participating in technological-based intensive ABA interventions, in both communication and socialization areas (Artoni et al., 2018).

The features of the BEHAVE application overcame some of the known limitations of the use of computation technologies in ABA. Many technologies do not permit monitoring the child's activities at home and are not fully customizable, forcing users to use activities or stimuli that are already in the system (Trevisan et al., 2019; Lopez-Herrejon et al., 2020). The possibility of remote monitoring may constitute added value at a time when the COVID-19 pandemic has limited face-to-face meetings, including concerning activities for the management of neurodevelopmental disorders. Moreover, the flexibility of the measure creation of the BEHAVE application partially fulfills this need even if other features could make it more flexible in future releases of the system (e.g., the possibility to observe more than one behavior at a time). The fact that the BEHAVE application provides users with automatic statistical analyses does not imply that the application can be used without specific training or that the decisions about behavioral interventions can be completely delegated to algorithms. In this regard, the BEHAVE project produced many educational contents and was carried out both face-to-face (in five European countries) and virtual training attended by hundreds of teachers from a lifelong learning perspective. Technological tools supporting the monitoring of behaviors must foster ethics (Mittelstadt et al., 2016), transparency and integrity of data and processes, and

promote awareness, knowledge, and collaboration between practitioners and caregivers to apply the best multidisciplinary treatment plan (Guinchat et al., 2020), maximizing children participation in a typical home and school setting, reducing the impact of their symptoms but also facilitating the management of the disorder by various educational agents involved in their education.

7. Conclusion

The results of the study showed that the BEHAVE application can be a useful tool supporting the data collection, monitoring, and analysis of behavioral interventions through SCEDs. The paper presented two cases of children with NDDs and severe language improvements that improved their mand repertoire through an ABA intervention and generalized these results also in the home and education environment.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://figshare.com/articles/dataset/data_CSV/19753639; <https://doi.org/10.6084/m9.figshare.19753639.v1>.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

GM, AC, NS, and SP contributed to the formulation of the idea and the design of the study. GM, GC, and DT developed the software. NF, SP, and LP carried out the interventions and gathered the clinical data. NS performed the statistical analyses. GM, AC, GC, DT, LS, NS, and SP wrote the first draft of the manuscript. PD reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

References

- Abel, M. H., and Sewell, J. (1999). Stress and burnout in rural and urban secondary school teachers. *J. Educ. Res.* 92, 287–293. doi: 10.1080/00220679909597608
- Aktaş, B., and Ciftci-Tekinarslan, I. (2018). The effectiveness of parent training a mothers of children with autism use of mand model techniques. *Int. J. Early Childhood Spec. Educ.* 10, 106–120. doi: 10.20489/intjecse.506861
- Allison, D. B., and Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: the case of the single case*. *Behav. Res. Ther.* 31, 621–631. doi: 10.1016/0005-7967(93)90115-B
- American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 5th Edn.* Washington, DC: Autor.
- Amstad, M., and Müller, C. M. (2020). Students' problem behaviors as sources of teacher stress in special needs schools for individuals with intellectual disabilities. *Front. Educ.* 4:159. doi: 10.3389/educ.2019.00159
- Artoni, S., Bastiani, L., Buzzi, M. C., Buzzi, M., Curzio, O., Pelagatti, S., et al. (2018). Technology-enhanced ABA intervention in children with autism: a pilot study. *Univ. Access Inf. Soc.* 17, 191–210. doi: 10.1007/s10209-017-0536-x
- Balogh, R., Bretherton, K., Whibley, S., Berney, T., Graham, S., Richold, P., et al. (2001). Sexual abuse in children and adolescents with intellectual disability. *J. Intellect. Disabil. Res.* 45, 194–201. doi: 10.1046/j.1365-2788.2001.00293.x
- Barnard-Brak, L., Richman, D. M., and Watkins, L. (2022). Introduction to the special section: translating advanced quantitative techniques for single-case experimental design data. *Perspect. Behav. Sci.* 45, 1–4. doi: 10.1007/s40614-022-00327-0
- Bentley, K., Kleiman, E., Elliott, G., Huffman, J., and Nock, M. (2018). Real-time monitoring technology in single-case experimental design research: opportunities and challenges. *Behav. Res. Ther.* 117, 87–96. doi: 10.1016/j.brat.2018.11.017
- Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory Psychol.* 24, 256–277. doi: 10.1177/0959354314525282
- Cafiero, J. M. (2009). *Comunicazione Aumentativa e Alternativa. Strumenti e Strategie per l'autismo e i Deficit di Comunicazione.* Trento: Edizioni Erickson.
- Camargo, S. P. H., Rispoli, M., Ganz, J., Hong, E. R., Davis, H., and Mason, R. (2014). A review of the quality of behaviorally-based intervention research to improve social interaction skills of children with ASD in inclusive settings. *J. Autism Dev. Disord.* 44, 2096–2116. doi: 10.1007/s10803-014-2060-7
- Cannella-Malone, H. I., Dueker, S. A., Barczak, M. A., and Brock, M. E. (2021). Teaching academic skills to students with significant intellectual disabilities: a systematic review of the single-case design literature. *J. Intellect. Disabil.* 25, 387–404. doi: 10.1177/1744629519895387
- Carbone, V. J., Sweeney-Kerwin, E. J., Attanasio, V., and Kasper, T. (2010). Increasing the vocal responses of children with autism and developmental disabilities using manual sign mand training and prompt delay. *J. Appl. Behav. Anal.* 43, 705–709. doi: 10.1901/jaba.2010.43-705
- Chiazze, G., Mariscalco, E., Chifari, A., Merlo, G., Goei, S. L., Mangina, E., et al. (2019). "The BASE system: a digital behavioral assessment tool for school environment" in Proceedings of *EdMedia+ Innovate Learning*. ed. J. T. Theo Bastiaens (Waynesville, NC: Association for the Advancement of Computing in Education (AACE)), 349–354.
- Cook, B. G., and Cook, S. C. (2013). Unraveling evidence-based practices in special education. *J. Spec. Educ.* 47, 71–82. doi: 10.1177/0022466911420877
- Cook, B. G., and Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Except. Child.* 79, 135–144. doi: 10.1177/0014402913079002021
- Cooper, J. O., Heron, T. E., and Heward, W. L. (2007). *Applied Behavior Analysis. (2nd Edn.)*. Upper Saddle River, NJ: Pearson Education.
- Costanzo, V., Chericoni, N., Amendola, F. A., Casula, L., Muratori, F., Scattoni, M. L., et al. (2015). Early detection of autism spectrum disorders: from retrospective home video studies to prospective 'high risk/sibling studies. *Neurosci. Biobehav. Rev.* 55, 627–635. doi: 10.1016/j.neubiorev.2015.06.006
- Daniels, A. M., Halladay, A. K., Shih, A., Elder, L. M., and Dawson, G. (2014). Approaches to enhancing the early detection of autism spectrum disorders: a systematic review of the literature. *J. Am. Acad. Child Adolesc. Psychiatry* 53, 141–152. doi: 10.1016/j.jaac.2013.11.002
- Dykens, E. M. (2015). Family adjustment and interventions in neurodevelopmental disorders. *Curr. Opin. Psychiatry* 28, 121–126. doi: 10.1097/YCO.0000000000000129
- Elsabbagh, M., Divan, G., Koh, Y.-J., Kim, Y. S., Kauchali, S., Marcin, C., et al. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Res.* 5, 160–179. doi: 10.1002/aur.239
- Evans, J. G. (1995). Evidence-based and evidence-biased medicine. *Age Ageing* 24, 461–463. doi: 10.1093/ageing/24.6.461
- Ghani, M. Z., Ahmad, A. C., and Ibrahim, S. (2014). Stress among special education teachers in Malaysia. *Procedia Soc. Behav. Sci.* 114, 4–13. doi: 10.1016/j.sbspro.2013.12.648
- Guinchat, V., Cravero, C., Lefèvre-Utile, J., and Cohen, D. (2020). Multidisciplinary treatment plan for challenging behaviors in neurodevelopmental disorders. *Handb. Clin. Neurol.* 174, 301–321. doi: 10.1016/B978-0-444-64148-9.00022-3
- Guo, H., Wang, T., Wu, H., Long, M., Coe, B. P., Li, H., et al. (2018). Inherited and multiple de novo mutations in autism/developmental delay risk genes suggest a multifactorial model. *Mol. Autism* 9, 1–12. doi: 10.1186/s13229-018-0247-z
- Hetzroni, O. E., and Tannous, J. (2004). Effects of a computer-based intervention program on the communicative functions of children with autism. *J. Autism Dev. Disord.* 34, 95–113. doi: 10.1023/B:JADD.0000022602.40506.bf
- Ho, H., Perry, A., and Koudys, J. (2021). A systematic review of behaviour analytic interventions for young children with intellectual disabilities. *J. Intellect. Disabil. Res.* 65, 11–31. doi: 10.1111/jir.12780
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., and Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Except. Child.* 71, 165–179. doi: 10.1177/001440290507100203
- Jacobson, N. S., and Christensen, A. (1996). Studying the effectiveness of psychotherapy: how well can clinical trials do the job? *Am. Psychol.* 51, 1031–1039. doi: 10.1037/0003-066X.51.10.1031
- Jawaid, A., Riby, D., Owens, J., White, S., Tarar, T., and Schulz, P. (2012). 'Too withdrawn' or 'too friendly': considering social vulnerability in two neurodevelopmental disorders. *J. Intellect. Disabil. Res.* 56, 335–350. doi: 10.1111/j.1365-2788.2011.01452.x
- Jurgens, A., Anderson, A., and Moore, D. W. (2009). The effect of teaching PECS to a child with autism on verbal behaviour, play, and social functioning. *Behav. Chang.* 26, 66–81. doi: 10.1375/bech.26.1.66
- Kamara, D., and Beauchaine, T. P. (2020). A review of sleep disturbances among infants and children with neurodevelopmental disorders. *Rev. J. Autism Dev. Disord.* 7, 278–294. doi: 10.1007/s40489-019-00193-8
- Karst, J. S., and Van Hecke, A. V. (2012). Parent and family impact of autism spectrum disorders: a review and proposed model for intervention evaluation. *Clin. Child. Fam. Psychol. Rev.* 15, 247–277. doi: 10.1007/s10567-012-0119-6
- Kazdin, A. E. (2021). Single-case experimental designs: characteristics, changes, and challenges. *J. Exp. Anal. Behav.* 115, 56–85. doi: 10.1002/jeab.638
- Lopez-Herrejon, R. E., Poddar, O., Herrera, G., and Sevilla, J. (2020). Customization support in computer-based technologies for autism: a systematic mapping study. *Int. J. Hum. Comput. Inter.* 36, 1273–1290. doi: 10.1080/10447318.2020.1731673
- Lundström, S., Forsman, M., Larsson, H., Kerekes, N., Serlachius, E., Långström, N., et al. (2014). Childhood neurodevelopmental disorders and violent criminality: a sibling control study. *J. Autism Dev. Disord.* 44, 2707–2716. doi: 10.1007/s10803-013-1873-0
- Lunsky, Y. (2004). Suicidality in a clinical and community sample of adults with mental retardation. *Res. Dev. Disabil.* 25, 231–243. doi: 10.1016/j.ridd.2003.06.004
- Makrygianni, M. K., Gena, A., Katoudi, S., and Galanis, P. (2018). The effectiveness of applied behavior analytic interventions for children with autism spectrum disorder: a meta-analytic study. *Res. Autism Spectr. Disord.* 51, 18–31. doi: 10.1016/j.rasd.2018.03.006
- Martin, G., and Pear, J. (2019). *Behavior Modification: What It Is and How To Do It*. New York: Routledge.
- Matson, J. L., and Kozlowski, A. M. (2011). The increasing prevalence of autism spectrum disorders. *Res. Autism Spectr. Disord.* 5, 418–425. doi: 10.1016/j.rasd.2010.06.004

- Meadan, H., Halle, J. W., and Ebata, A. T. (2010). Families with children who have autism spectrum disorders: stress and support. *Except. Child.* 77, 7–36. doi: 10.1177/001440291007700101
- Merlo, G., Chifari, A., Chiazzeze, G., Taibi, D., Alves, S., McGee, C., et al. (2020). “Introducing evidence-based practices to manage problem behaviours at school: the BEHAVE application” in *European Conference on E-Learning*. ed. C. Busch (Sonning Common, UK: Academic Conferences International Limited), 335–XVII.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data Soc.* 3:205395171667967. doi: 10.1177/2053951716679679
- Moderato, P., and Copelli, C. (2010). L'analisi comportamentale applicata: seconda parte: Metodi e procedure. *Autismo e Disturbi dello Sviluppo* 8, 191–233.
- Natesan Batley, P., Contractor, A. A., and Caldas, S. V. (2020). Bayesian time-series models in single case experimental designs: a tutorial for trauma researchers. *J. Trauma. Stress.* 33, 1144–1153. doi: 10.1002/jts.22614
- Norbury, C. F., and Paul, R. (2015). “Disorders of speech, language, and communication” in *Rutter's Child and Adolescent Psychiatry - 6th edn.* eds. A. Thapar, D. Pine, J. Leckman, S. Scott, M. Snowling, and E. Taylor (New York: Wiley-Blackwell), 683–701.
- Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., and Harris, K. R. (2005). Research in special education: scientific methods and evidence-based practices. *Except. Child.* 71, 137–148. doi: 10.1177/001440290507100201
- Parker, R. I., Vannest, K. J., Davis, J. L., and Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: tau-u. *Behav. Ther.* 42, 284–299. doi: 10.1016/j.beth.2010.08.006
- Peters, T., and Gillberg, C. (1999). *Autism: Medical and Educational Aspects*. London: John Wiley & Sons Incorporated.
- Perone, M. (2019). How I learned to stop worrying and love replication failures. *Perspect. Behav. Sci.* 42, 91–108. doi: 10.1007/s40614-018-0153-x
- Peters-Scheffer, N., Didden, R., Korzilius, H., and Sturmey, P. (2011). A meta-analytic study on the effectiveness of comprehensive ABA-based early intervention programs for children with autism spectrum disorders. *Res. Autism Spectr. Disord.* 5, 60–69. doi: 10.1016/j.rasd.2010.03.011
- Presti, G., Moderato, P., Gentile, R., and Chase, P. (2002). Le relazioni verbali: Analisi e caratteristiche funzionali. P. Moderato, G. Presti and PN Chase (a cura di), *Pensieri, Parole e Comportamento. Un'analisi Funzionale Delle Relazioni Linguistiche*, Milano, McGraw-Hill.
- Reichow, B., Barton, E. E., Boyd, B. A., and Hume, K. (2018). Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders (ASD). *Cochrane Database Syst. Rev.* 5. doi: 10.1002/14651858.CD009260.pub3
- Reiter, S., Bryen, D. N., and Shachar, I. (2007). Adolescents with intellectual disabilities as victims of abuse. *J. Intellect. Disabil.* 11, 371–387. doi: 10.1177/1744629507084602
- Richards, A. R., Hemphill, M. A., and Templin, T. J. (2018). Personal and contextual factors related to teachers' experience with stress and burnout. *Teach. Teach.* 24, 768–787. doi: 10.1080/13540602.2018.1476337
- Ross, D. E., and Greer, R. D. (2003). Generalized imitation and the mand: inducing first instances of speech in young children with autism. *Res. Dev. Disabil.* 24, 58–74. doi: 10.1016/S0891-4222(02)00167-1
- Saini, V., and Roane, H. S. (2018). Technological advances in the experimental analysis of human behavior. *Behav. Anal. Res. Pract.* 18, 288–304. doi: 10.1037/bar0000124
- Schlier, B., Krkovic, K., Clamor, A., and Lincoln, T. M. (2019). Autonomic arousal during psychosis spectrum experiences: results from a high resolution ambulatory assessment study over the course of symptom on- and offset. *Schizophr. Res.* 212, 163–170. doi: 10.1016/j.schres.2019.07.046
- Shapiro, M. (1966). Generality of psychological processes and specificity of outcomes. *Percept. Mot. Skills* 23:16. doi: 10.2466/pms.1966.23.1.16
- Shiffman, S., Stone, A. A., and Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4, 1–32. doi: 10.1146/annurev.clinpsy.3.022806.091415
- Shingleton, R. M., Pratt, E. M., Gorman, B., Barlow, D. H., Palfai, T. P., and Thompson-Brenner, H. (2016). Motivational text message intervention for eating disorders: a single-case alternating treatment design using ecological momentary assessment. *Behav. Ther.* 47, 325–338. doi: 10.1016/j.beth.2016.01.005
- Sidman, M. (1952). A note on functional relations obtained from group data. *Psychol. Bull.* 49, 263–269. doi: 10.1037/h0063643
- Sidman, M. (1960). *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*. New York: Basic Books.
- Simonoff, E. (2015). “Intellectual disability” in *Rutter's Child and Adolescent Psychiatry - 6th edn.* eds. A. Thapar, D. Pine, J. Leckman, S. Scott, M. Snowling, and E. Taylor (New York: Wiley-Blackwell), 719–737.
- Skinner, B. (1938). *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century-Crofts.
- Skinner, B. F. (1956). A case history in scientific method. *Am. Psychol.* 11, 221–233. doi: 10.1037/h0047662
- Skinner, B. F. (1957). *Verbal Behavior*. New York: Appleton-Century-Crofts.
- Smith, J. D. (2012). Single-case experimental designs: a systematic review of published research and current standards. *Psychol. Methods* 17, 510–550. doi: 10.1037/a0029312
- Smith, T. E., Thompson, A. M., and Maynard, B. R. (2022). Self-management interventions for reducing challenging behaviors among school-age students: a systematic review. *Campbell Syst. Rev.* 18:e1223. doi: 10.1002/cl2.1223
- Spachos, D., Chifari, A., Chiazzeze, G., Merlo, G., Doherty, G., and Bamidis, P. (2014). WHAAM: A Mobile Application for Ubiquitous Monitoring of ADHD Behaviors. In 2014 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL2014) (IEEE), 305–309.
- Sundberg, M. L. (2008). *VB-MAPP Verbal Behavior Milestones Assessment and Placement Program: A Language and Social Skills Assessment Program for Children with Autism or Other Developmental Disabilities: Guide*. Concord, CA: Mark Sundberg.
- Sundberg, M. L., and Michael, J. (2001). The benefits of skinner's analysis of verbal behavior for children with autism. *Behav. Modif.* 25, 698–724. doi: 10.1177/0145445501255003
- Sundberg, M. L., and Sundberg, C. A. (2011). Intraverbal behavior and verbal conditional discriminations in typically developing children and children with autism. *Anal. Verbal Behav.* 27, 23–44. doi: 10.1007/BF03393090
- Tate, R. L., and Perdices, M. (2019). *Single-Case Experimental Designs for Clinical Research and Neurorehabilitation Settings: Planning, Conduct, Analysis and Reporting*. London and New York: Routledge.
- Tosto, C., Hasegawa, T., Mangina, E., Chifari, A., Treacy, R., Merlo, G., et al. (2021). Exploring the effect of an augmented reality literacy programme for reading and spelling difficulties for children diagnosed with ADHD. *Virtual Reality* 25, 879–894. doi: 10.1007/s10055-020-00485-z
- Trevisan, D. F., Becerra, L., Benitez, P., Higbee, T. S., and Gois, J. P. (2019). A review of the use of computational technology in applied behavior analysis. *Adapt. Behav.* 27, 183–196. doi: 10.1177/1059712319839386
- Valentine, A. Z., Brown, B. J., Groom, M. J., Young, E., Hollis, C., and Hall, C. L. (2020). A systematic review evaluating the implementation of technologies to assess, monitor and treat neurodevelopmental disorders: a map of the current evidence. *Clin. Psychol. Rev.* 80:101870. doi: 10.1016/j.cpr.2020.101870
- Vlaeyen, J. W., Wicksell, R. K., Simons, L. E., Gentili, C., De, T. K., Tate, R. L., et al. (2020). From boulder to Stockholm in 70 years: single case experimental designs in clinical research. *Psychol. Rec.* 70, 659–670. doi: 10.1007/s40732-020-00402-5
- Wachtel, P. L. (2010). Beyond “ESTs”: problematic assumptions in the pursuit of evidence-based practice. *Psychoanal. Psychol.* 27, 251–272. doi: 10.1037/a0020532
- Westen, D., and Bradley, R. (2005). Empirically supported complexity: rethinking evidence-based practice in psychotherapy. *Curr. Dir. Psychol. Sci.* 14, 266–271. doi: 10.1111/j.0963-7214.2005.00378.x
- Wichers, M., and Groot, P. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychother. Psychosom.* 85, 114–116. doi: 10.1159/000441458
- Williams, J. G., Higgins, J. P., and Brayne, C. E. (2006). Systematic review of prevalence studies of autism spectrum disorders. *Arch. Dis. Child.* 91, 8–15. doi: 10.1136/adc.2004.062083
- Yu, Q., Li, E., Li, L., and Liang, W. (2020). Efficacy of interventions based on applied behavior analysis for autism spectrum disorder: a meta-analysis. *Psychiatry Investig.* 17, 432–443. doi: 10.30773/pi.2019.0229
- Zheng, Z. K., Staubitz, J., Jessel, J., Fruchtmann, T., and Sarkar, N. (2022). Validating a computerized program for supporting visual analysis during functional analysis: the problem behavior multilevel interpreter (PB. MI). *Behav. Anal. Pract.* 15, 485–494. doi: 10.1007/s40617-021-00656-7

Frontiers in Education

Explores education and its importance for individuals and society

A multidisciplinary journal that explores research-based approaches to education for human development. It focuses on the global challenges and opportunities education faces, ultimately aiming to improve educational outcomes.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact



Frontiers in Education

