frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 28 March 2023
DOI 10.3389/feduc.2023.1066837

OPEN ACCESS

EDITED BY
Yong Luo,
NWEA,
United States

REVIEWED BY
Yu Bao,
James Madison University,
United States
Fei Zhao,
NWEA,
United States

*CORRESPONDENCE
Rielke Bogaert
✉ Rielke.Bogaert@ugent.be

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 11 October 2022
ACCEPTED 01 March 2023
PUBLISHED 28 March 2023

CITATION
Bogaert R, Merchie E, Aesaert K and Van Keer H (2023) The development of the reading comprehension—Progress monitoring (RC-PM) tool for late elementary students. *Front. Educ.* 8:1066837. doi: 10.3389/feduc.2023.1066837

The development of the reading comprehension—Progress monitoring (RC-PM) tool for late elementary students

Rielke Bogaert[1]*, Emmelien Merchie[1], Koen Aesaert[2] and Hilde Van Keer[1]

[1]Department of Educational Studies, Ghent University, Ghent, Belgium, [2]Educational Effectiveness and Evaluation, KU Leuven, Leuven, Belgium

Notwithstanding reading comprehension is a key competence in today's society, many late elementary students struggle with it. In this respect, effective instructional incentives are required to foster students' reading comprehension. However, appropriate assessment instruments to monitor students' reading comprehension on a regular basis and to make substantiated instructional decisions are lacking. Therefore, a Reading Comprehension – Progress Monitoring tool was developed, consisting of six parallel tests equivalent in difficulty and length. To this aim, classical test theory analyses, item response theory analyses, and automated test assembly were conducted (n=3,269 students). Suggestions for future research and practice are discussed.

KEYWORDS

reading comprehension, late elementary education, progress monitoring assessment, test development, item response theory

1. Introduction

Reading comprehension is key to students' success in school, as well as in life (Wijekumar et al., 2019). Especially in the critical period of late elementary education, developing appropriate skills to comprehend expository texts becomes increasingly important (Keresteš et al., 2019). During this period, students shift from learning to read, to reading to learn. In this respect, they are increasingly expected to read, process, and comprehend expository text information independently (Meneghetti et al., 2007). Unfortunately, many late elementary school children struggle with reading comprehension, especially comprehension of expository texts (Rasinski, 2017). In order to stimulate and foster elementary students' reading comprehension of expository texts, substantiated and effective instructional practices are required. More specifically, teachers need appropriate assessment tools to make substantiated and systematical instructional decisions in order to stimulate high levels of performance among all students (Stecker et al., 2008; Zeuch et al., 2017). In order to make these decisions, the importance of monitoring the reading progress of students on a regular basis has received growing attention in reading research in the past two decades (Förster and Souvignier, 2011; Fuchs, 2017). Moreover, research indicated that monitoring students' progress to adjust and evaluate teachers' instruction can result in significantly higher achievement gains for reading comprehension (Stecker et al., 2005; Förster and Souvignier, 2014). However, appropriate expository text comprehension progress monitoring tools enabling administration on a regular basis are lacking. Most of the progress monitoring tools focus on reading fluency (e.g., CBM-R; Ardoin et al., 2013) instead of on reading comprehension. Moreover, when there is a focus on

reading comprehension, the monitoring instruments often lack extensive academic empirical and theoretical underpinning (Leslie and Caldwell, 2014).

Therefore, the purpose of the current study is to develop a Reading Comprehension Progress Monitoring (RC-PM) tool for late elementary school students, focusing on expository texts. Since text comprehension of expository texts is the focus of this study, the use of the term 'reading comprehension' further in the manuscript should be read and interpreted as 'reading comprehension of expository texts'. In what follows, first the theoretical background of this study is described. Next, an overview is provided on what constitutes a high-quality assessment instrument regarding reading comprehension. In this respect, the benefits and drawbacks of already existing assessment instruments in literature are reviewed. Finally, the specific assessment form of 'progress monitoring' will be outlined in depth.

2. Reading comprehension

2.1. Theoretical background

Definitions of reading comprehension often stress the complex and multifaceted nature of it (e.g., van Dijk and Kintsch, 1983; Snow, 2002; Cain et al., 2004; Randi et al., 2005; Kendeou et al., 2014; Castles et al., 2018; Follmer and Sperling, 2018). More specifically, the reading comprehension process is affected by an interplay of various factors and mental activities (e.g., reader and text characteristics, the socio-cultural context; Merchie et al., 2019). In sum, reading comprehension is a complex, multifaceted process of extracting and creating meaning from what is read (van Dijk and Kintsch, 1983; Snow, 2002; Randi et al., 2005; Stuart et al., 2008; Castles et al., 2018). Reading comprehension is strongly related with reading fluency within the early stages of reading development (Lee and Chen, 2019; Torppa et al., 2020). More specifically, students learn in the early reading stages to decode words, letters, and phrases accurately and fluently (i.e., reading fluency) and learn to comprehend the meaning of the written text (i.e., reading comprehension; Harlaar et al., 2007; Lee and Chen, 2019). However, reading comprehension is considered as the ultimate reading goal, since it becomes increasingly important throughout students' school career (Keresteš et al., 2019; Lee and Chen, 2019).

Throughout the years, various theoretical models of reading comprehension were developed stressing either bottom-up processes in reading comprehension (e.g., decoding), top-down processes (e.g., taking into account prior knowledge), or both (i.e., interactive models; Houtveen et al., 2019). The present study builds upon the construction-integration (CI)-model of Kintsch (1998, 2005) to define and operationalize reading comprehension since this model is considered as the most comprehensive interactive model (McNamara and Magliano, 2009). Moreover, this framework has already strong empirical backing in earlier studies (e.g., Elleman and Oslund, 2019; Stevens et al., 2019). In this respect, the CI model of Kintsch (1998, 2005) is still very valuable to operationalize reading comprehension and considers its complex nature. According to the CI model, good comprehenders master three levels of comprehension: (1) surface model, (2) textbase model, and (3) situation model. The surface model consists of the texts' literal representation, the words and phrases themselves. At the textbase level, the meaning of the text is represented as a network of propositions and concepts from the text at micro and macro level. In the situation model, information of the text is integrated with prior knowledge/experiences.

2.2. Assessment

A valid assessment of reading comprehension of expository texts is complicated and challenging due to the complex nature described above (Collins et al., 2018; Calet et al., 2020). Therefore, the question arises what constitutes a high-quality, valid, and reliable assessment instrument for reading comprehension. A wide diversity of instruments has already been described in the literature. These instruments differ according to the (a) applied response formats, (b) question types, and (c) the assessment frequency. In what follows, an overview of these differences in assessment instruments is provided, together with some critical considerations.

2.2.1. Response formats

Reading comprehension has been assessed in prior research by means of different response formats (i.e., free-text recall, true/false statements, cloze tasks, open-ended, and multiple-choice questions; Collins et al., 2018).

Free-text recall consists of retelling what the reader has read (e.g., used in the study of Roberts et al., 2005). This type of assessment, however, focusses more on how well a reader can reproduce or remember text information instead of on how well the text is understood (Barbe, 1958; Collins et al., 2018).

Further, sentence verification tasks and cloze tasks mainly focus on assessing sentence-level understanding (Collins et al., 2018). Sentence verification tasks are for instance used as true/false tests wherein students must verify the accuracy of a statement in the original text (Collins et al., 2018). Cloze tasks consist of an incomplete text whereby students must restore systematically placed missing words (Jensen and Elbro, 2022). However, the validity of assessing reading comprehension by means of this specific response format is often criticized (Muijselaar et al., 2017; Jensen and Elbro, 2022). For example, some researchers conclude that cloze tasks are more sensitive to reading speed and word level processes than higher-order comprehension processes (Keenan et al., 2008; Muijselaar et al., 2017). Therefore, the content validity of this response format as indicator of students' comprehension (i.e., the extent to which the test content is representative for the skills or domain that the test aims to assess) can be questioned (Hoover et al., 2003).

Open-ended and multiple-choice questions have shown to carry a greater potential in assessing more higher-cognitive processes, requiring the construction of a situation model (Collins et al., 2018). Open-ended or constructed-response questions, consisting of an open answer format, allow participants to answer in a free and individual way. However, this free format requires high output demands in terms of linguistic skills to formulate appropriate responses (Weigle et al., 2013; Calet et al., 2020). Multiple-choice (MC) questions, introduced by Davis (1944), provides participants with predefined answer options (Leslie and Caldwell, 2014; Collins et al., 2018), which allows easy and quick test administration in large groups (Calet et al., 2020; Nundy et al., 2022). MC questions can be used to assess aspects of students' performance in an effective and reliable way (Brady, 2005; Nundy

et al., 2022). Moreover, MC items require fewer writing competences of students than open-ended questions (Weigle et al., 2013; Green, 2017). To avoid interdependency with the passage content, researchers have pointed out the importance of including longer and various passages in reading comprehension tests (Leslie and Caldwell, 2014). Since longer passages provides more context to students, this limits the influence of students' decoding skills as well (Calet et al., 2020). However, researchers also point at the high processing demands of answering MC questions, since readers have to weigh several answer options against each other (Calet et al., 2020). In this respect, it is advised to include MC items with three answer options, since these maintain psychometric quality if a sufficient number of items are included (i.e., item difficulty and discrimination does almost not reduce from four to three options items) and maximizes efficiency by reducing the cognitive load and test administration time (Rodriguez, 2005; Haladyna and Rodriguez, 2013; Simms et al., 2019; Holzknecht et al., 2021; Sarac and Feinberg, 2022). Also for students with learning/reading difficulties, three option MC items are recommended (Goegan et al., 2018).

2.2.2. Question types

Besides the response format, researchers draw upon the importance of considering the question type when developing appropriate assessment instruments for reading comprehension (Calet et al., 2020). Rather than focusing exclusively on quantity questions (e.g., how many, how much), which refer exclusively to reading comprehension on the surface model of Kintsch (1998, 2005), it is recommended to include different question types in view of assessing the complex comprehension process reflected in the CI model as accurate as possible (Barbe, 1958; Calet et al., 2020). In this respect, both current international (e.g., the Progress in International Reading Literacy Study; PIRLS) and national (e.g., Norwegian Reading Tests; NRT) tests include questions referring to various comprehension processes (i.e., retrieve, interpret, and evaluate text information; Støle et al., 2020).

2.2.3. Assessment frequency

Previous research also points at the importance of the test administration frequency. In this respect, monitoring students' reading comprehension on a regular basis to make substantiated and systematical instructional decisions is receiving increasing attention in reading research (Förster and Souvignier, 2011; Fuchs, 2017). Assessing students' reading comprehension at several points in time allows for monitoring students' progress, detecting comprehension difficulties, adjusting instruction accordingly, and evaluating the effectiveness of the instruction (Förster and Souvignier, 2015; Calet et al., 2020). This will be outlined in-depth in the following section.

2.3. Progress monitoring assessment

As described above, monitoring the progress of students' reading comprehension on a regular basis is found to be very important since it can contribute to identifying struggling students and provides teachers with information in view of optimizing students' learning process as well as their own instruction (Zeuch et al., 2017; Förster et al., 2018). Progress monitoring assessments are characterized by assessment at regular fixed intervals and a time efficient administration process (Peters et al., 2021). A minimum of at least two tests is recommended when using reading comprehension tests for the above-mentioned educational purposes (Calet et al., 2020). Two frequently discussed types of progress monitoring assessments in the literature are Learning progress assessment (LPA) and curriculum-based measurement (CBM). LPA focuses on monitoring all students learning process, while CBM (Deno, 1985) was introduced to monitor the learning progress of poorly achieving students (Zeuch et al., 2017; Förster et al., 2018).

The targeted use of progress monitoring assessments such as CBM and LPA in order to make substantiated and systematically instructional decisions has produced repeatedly proven effects in fostering student learning, particularly students' reading comprehension progress (Förster and Souvignier, 2014, 2015; Förster et al., 2018). However, despite the importance of assessing reading comprehension at regular fixed intervals, most progress monitoring tools have focused on reading fluency (e.g., Oral Reading Fluency CBM-Passage Fluency (ORF-PF); Tolar et al., 2014; Curriculum-Based Measurement in Reading (CBM-R); Ardoin et al., 2013) instead of on comprehension. As to the researchers that have included reading comprehension progress monitoring tools in their studies, three shortcomings can be noticed. First, although some researchers explicitly indicated that they have included a progress monitoring tool for reading comprehension in their study, they actually applied the criticized cloze task (e.g., AIMSweb Maze CBM Reading Comprehension subtest (AIMS); Shinn and Shinn, 2002; Tolar et al., 2014). Second, progress instruments that do enable reading comprehension administration on a regular basis, mainly lack an extensive academic empirical and theoretical underpinning (Leslie and Caldwell, 2014). Third, even if the reading comprehension test was empirically and theoretically grounded (e.g., based on the comprehension model of Kintsch, 1998, 2005), questions referred to a specific text passage (e.g., Förster et al., 2018) although integrating various passages is advised in literature (Leslie and Caldwell, 2014).

3. Objective of the study

Although some reading comprehension tests exists (e.g., CITO Reading Comprehension Test, Cito, 2014; Diatekst, Hacquebord et al., 2005), there is currently no high-quality, valid, and reliable measurement instrument to monitor students' expository reading comprehension skills on a regular basis. More specifically, the majority of the currently available progress monitoring instruments put more emphasis on reading fluency than on reading comprehension (e.g., CBM-R; Ardoin et al., 2013). Moreover, the instruments wherein reading comprehension is emphasized, frequently lack overall academic empirical and theoretical support (Leslie and Caldwell, 2014). In this respect, there is an urgent need for a theoretically underpinned and empirically sound reading comprehension progress monitoring tool aligned to the crucial period of upper elementary education.

Based on the literature overview, three important recommendations can be distilled as to what constitutes a high-quality, valid, and reliable assessment instrument to monitor students' reading comprehension of expository texts. First, as to the response format, it is advised to include MC questions with three-option answers referring to various multiple-paragraph texts. This allows to efficiently map the reading comprehension performance of large student groups and limits the influence decoding skills

might play (Calet et al., 2020). Second, a wide range of question types - aiming to assess text comprehension at various levels - should be included (Barbe, 1958; Calet et al., 2020). Third, as to the frequency of reading comprehension assessment, regular progress monitoring (e.g., several times a year) is required to underpin substantiated and effective instructional decisions (Förster and Souvignier, 2015; Förster et al., 2018). Despite some reading comprehension students monitoring systems are available, instruments covering all these evidence-informed recommendations are currently lacking (i.e., focus on expository texts, including a wide range of MC questions and suitable to be taken easily at a regular base).

Therefore, the aim of this study was to develop a reading comprehension progress monitoring tool for late elementary students, taken all these recommendations into account.

4. Materials and methods

4.1. Test development process

The test development was guided by the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). Below, the test development process is described chronologically throughout sixth consecutive steps (i.e., construct development, item development, item review, creating test forms, pilot study, and large-scale study). Within these sixth steps, test specifications derived from the Standards are integrated (i.e., specifications regarding content, format, length, administration, scoring, and psychometrics).

4.1.1. Step 1: Construct development

In the first step, the construct of expository text comprehension was clearly defined (see the theoretical framework of this study) based on the CI model of Kintsch (1998, 2005). More particularly, according to the CI model, reading comprehension is realized by constructing literal mental representations of the meaning of texts (i.e., surface model), by interpreting texts' propositions at micro and macro level (i.e., textbase model), and by integrating these representations with prior knowledge/experiences (i.e., situation model). Further, the content of the Flemish curriculum, the context of this study, on expository text comprehension was considered. The reading curriculum consists of ten minimum attainment goals for all students, including one explicitly focusing on comprehending expository texts (i.e., students can retrieve and organize expository text information).

4.1.2. Step 2: Item development

In a next step, texts were selected and items were developed. The initially developed reading comprehension progress monitoring tool (RC-PM tool) consisted of 60 expository texts with accompanying MC items.

The 60 expository texts selected for the initial pool had varying lengths (ranging between 81 and 506 words). Following the recommendations in prior research (see introduction), both one, two, and more paragraph texts were incorporated to limit the influence of students' decoding skills (Calet et al., 2020). Further, varied topics (i.e., food, plants and animals, arts, STEM, society, economics, language, sport) were covered in the selected texts as well. These topics represent different study fields in Flemish secondary education to address a range of student interests.

As to the test items, a template of possible questions per comprehension level was created (Table 1). The template was grounded in the CI model and in existing reading comprehension (test) manuals (e.g., Cito test, Being upside-down from reading [Onderste boven van lezen], Read yourself wiser: reading comprehension tests [Lees je wijzer: toetsen begrijpend lezen], Reading Journey (Leesreis), Easy Curriculum Based Measurement Reading). Eight educational researchers applied this template to create test items at the level of the surface, textbase, and situation model for the selected texts. Depending on the length of the text, 8 to 15 test items were developed per text. Table 2 presents some example items per comprehension level.

As to the answer format of the test items, three-option MC items were opted for based on the literature. More specifically, MC items allow more to probe higher-cognitive processes and are easy to administer (Collins et al., 2018; Calet et al., 2020). In addition, the high processing demand of MC options is countered by using three-option MC questions as recommended by the meta-analysis of Rodriguez (2005). The MC items belonging to one text passage, were arranged in the sequence they were addressed in the text. MC items focusing on the full text were placed last. To avoid logical patterns in the answer options, the three answer options were alphabetically ordered per MC item.

4.1.3. Step 3: Item review

After the initial text selection and item development, an expert panel was consulted. This panel consisted of three primary school teachers and three educational researchers on reading comprehension. They reviewed in detail the extent to which the topic and content of each text in the initial text pool was appropriate for the target group of late elementary students. Based on this feedback, less appropriate topics were removed (e.g., political entities, the #metoo movement against sexual inappropriate behavior). Next, the texts' Reading Fluency Level (RFL) and Reading Comprehension Level (RCL) were analyzed by the Dutch Central Institute for Test Development regarding (CITO). This institute operationalizes RFL by means of AVI levels (i.e., Analyse Van Individualiseringsvormen [Analysis of Individualization Forms]), consisting of twelve difficulty levels ranging from 'AVI-start' (i.e., the most basic level) to 'AVI-Plus' (i.e., the most difficult level). Comparably, RCL is operationalized by means of CLIB levels (index (i.e., Cito Lees Index voor het Basis- en speciaal onderwijs [Cito Reading Index for Elementary and Special Education])), consisting of eight levels ranging from 'CLIB-start' to 'CLIB-Plus'. Finally, all texts were professionally screened for spelling and language errors.

Parallel to the texts, the MC test items in the initial item pool were reviewed by a panel of experts to guarantee the quality of the test. More specifically, this expert panel consisted of ten late elementary teachers, one pedagogical counselor experienced in language skills of elementary students, and three educational researchers in reading comprehension. This expert panel reviewed the content validity (i.e., agreement between the assessed construct and the test content), the comprehensibility and clarity of the test items. As to the content validity, the pedagogical counselor and the educational experts criticized whether the distribution of the items

TABLE 1 Template of possible questions for each reading comprehension level.

	Surface model[1]	Textbase model[1]	Situation model[1]
Description	Texts' literal representation	Texts' propositions at micro and macro level	Integration of text information with prior knowledge/experiences
Template for the development of the test items	Meaning at micro level • What does '…' mean in the context of this text? • What is a(*n*) '…' in the context of this text? • Which sentence means (almost) the same as the sentence '…'?	Meaning at macro level • What does '…' mean in the context of this text? Referrals • To what does 'he'/'who'/… refer? • To what does 'that/there' refer?	Evaluate the relevance and reliability of arguments/evidence used in the text • Is this information '…' necessary/useful? • Is this information '…'/the text reliable?
	Synonyms/replacement • What other word can you use for the word '…'? • How could the writer have written this sentence as well? • In the text, another word is used for '…'. This word is:	Signal words • Which word is best to replace 'briefly' with? • Which word is best to put before the second sentence? (e.g., *and, so, but, because*) • Which linking phrase/connecting word can you use to indicate the connection between '…' and '…'?	Predicting • What is a logical additional paragraph of this text? • Which sentence fits the end of the text? • Suppose that … • Can you think of what … • What do you expect after the sentence '…'?
	Expression • What does the expression '…' mean in the context of this text? • In which sentence is figurative language used? • Which word from the text means the same as this expression?	Text structure • What do the writers do in this piece/part? (e.g., *summarize, give examples*) • What do these sentences have to do with each other? (e.g., *sentence 2 explains sentence 1; sentence 2 is the cause of sentence 1*) • What is the structure of the text? • Which of the following events occurred first? • The text is structured as follows: (e.g., *causes-solutions, causes-effects, information-solutions*)	Take a different view • What does the author want to say with (the statement) '…'? • What does the author claim in the text? • In what ways does the author's point of view influence the text content? • How do you think (child/mother) would react to/interpret this text? • What constitutes an argument against the text? • Suppose '…'. What argument would be no longer valid?
		Summarizing • What is the main idea of this text/paragraph? • What is a good alternative (sub)title for this text/paragraph? • What sentence best summarizes this piece of text? • What question does the text answer?	Text purpose • What is the authors' intent? • Where might you encounter this text? (e.g., *newspaper, novel*) • The purpose of this text is … (e.g., *to inform, persuade*) • For whom is this text important/interesting to read? • Who could have written this text?

[1]The three reading comprehension levels are based on the CI model of Kintsch (1998, 2005).

over the three comprehension levels was as evenly as possible. Furthermore, the comprehensibility and clarity of the test items was updated in various ways. For instance, in consultation with the teachers, the wording of the test items was aligned with the school language (e.g., 'section' was replaced by 'paragraph'; 'contrast' was replaced by 'difference'). When an item focuses on a specific word or sentence in the text, the line number was added in view of readability. Further, the wording of the questions was aligned across all items (i.e., using the same formulation for similar questions). Also, some test items were more concretized (e.g., 'What does the word '…' mean?' was replaced by 'What does the word '…' mean in the context of this text?'). Feedback was also provided on the answer options of the items (e.g., the options are too similar).

Based on the CITO analyses and the expert panel review, 35 of the initial 60 texts were retained, with 8 to 12 MC items per text (363 MC items in total). More specifically, texts that were too easy/difficult or too short/long were removed as well as test items with an insufficient quality (e.g., ambiguous or confusing items). These 35 texts, ranging from 95 to 387 words per text, contain 10 one paragraph texts

TABLE 2 Example items according to the three reading comprehension levels (translated from Dutch).

Reading comprehension level[1]	Description	Example items
Surface model	Texts' literal representation	*What does the word 'similar' mean in the context of this text?*
		A. Equivalent
		B. Identical
		C. Comparable
		What is a 'plantation' in the context of this text?
		A. An area where different types of plants grow.
		B. A greenhouse where trees or plants grow wild.
		C. A piece of land where people grow plants and trees.
Textbase model	Texts' propositions at micro and macro level	*To what does 'it' refer in the sentence '[…] because you push it aside.'?*
		A. The surface
		B. The water
		C. The water and the surface
		What is the relation between the following sentences?
		Sentence a: *France will ban smartphones in primary and secondary schools from September 2018.*
		Sentence b: *Many French teachers complained about smartphones in classes.*
		A. Sentence b is an example to sentence a.
		B. Sentence b is the cause of sentence a.
		C. Sentence b is the solution to sentence a.
Situation model	Integration of text information with prior knowledge/ experiences	*Which paragraph could you add at the end of the text?*
		A. A paragraph entitled 'How do fish get into the water'?
		B. A paragraph entitled 'Why do diving goggles always get stuck'?
		C. A paragraph entitled 'Why do we stay afloat when swimming in the sea'?
		Is the information in this text reliable?
		A. No, because the writer is telling his or her opinion.
		B. Yes, because the writer refers to several reports and researchers.
		C. No, because the text is written based on personal experiences.

[1]The three reading comprehension levels are based on the CI model of Kintsch (1998, 2005).

($M = 127.10$, SD = 23.28) and 25 two or more paragraph texts ($M = 262.76$, SD = 38.85). Table 3 represents the Reading Fluency Level and the Reading Comprehension Level of the 35 selected texts. In terms of content, the 363 MC items were approximately evenly distributed across the three comprehension levels (i.e., 121 surface, 123 textbase, and 119 situation level items).

4.1.4. Step 4: Creating test forms

In view of keeping the test administration time feasible and avoid fatigue, it was practically impossible to administrate 35 texts with each 8 to 12 MC items to every examinee. Therefore, a stepwise system was implemented wherein students received partially different reading comprehension tests (see Figure 1). More specifically, the first two reading texts of the test (i.e., a one paragraph text of 111 words and a four-paragraph text of 251 words) with 21 MC items in total were completed by all students, also called 'the anchor texts' during a first test moment (see further). These anchor texts are of importance to equate different test forms on one ability scale, making the scores of the different forms interchangeable (Dorans et al., 2010). Further, for the remaining 33 texts, the participants were allocated to 20 groups of each minimum 150 students. Supplementary to the anchor texts, each group completed three to four texts with each 8 to 12 MC items (i.e., 32 to 44 additional items in total per examinee) during a second test moment (see further). In this respect, the ratio of anchor

TABLE 3 Percentage of texts in each reading fluency and comprehension level.

Reading level	Description	Percentage of texts per level
Reading fluency level		
AVI M6	Expected reading fluency level in the middle of fourth grade	2.86%
AVI E6	Expected reading fluency level at the end of fourth grade	17.14%
AVI M7	Expected reading fluency level in the middle of fifth grade	14.29%
AVI E7	Expected reading fluency level at the end of fifth grade	11.43%
AVI PLUS	Expected reading fluency level in the sixth grade	54.28%
Reading comprehension level		
CLIB-7	Expected reading comprehension level in the fifth grade	25.72%
CLIB-8	Expected reading comprehension level in the sixth grade	57.14%
CLIB-PLUS	Expected reading comprehension level at the end of the sixth grade	17.14%

	Anchor texts		Short texts									Long texts																							
TEXT	Anchor text 1	Anchor text 2	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Text 9	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Text 9	Text 10	Text 11	Text 12	Text 13	Text 14	Text 15	Text 16	Text 17	Text 18	Text 19	Text 20	Text 21	Text 22	Text 23	Text 24
Group 1	x	x	x																			x									x				
Group 2	x	x	x										x																		x				
Group 3	x	x		x									x										x												
Group 4	x	x		x										x									x												x
Group 5	x	x			x									x										x									x		
Group 6	x	x			x										x									x										x	
Group 7	x	x				x									x										x									x	
Group 8	x	x				x										x									x										
Group 9	x	x					x									x										x									
Group 10	x	x					x										x									x									
Group 11	x	x						x									x										x								
Group 12	x	x						x										x									x								
Group 13	x	x							x									x										x							
Group 14	x	x							x										x									x							
Group 15	x	x								x									x										x						
Group 16	x	x								x										x									x						
Group 17	x	x									x									x										x		x			
Group 18	x	x									x										x									x					x
Group 19	x	x										x									x										x	x			
Group 20	x	x										x										x									x		x		

FIGURE 1
Stepwise system of the original test pool. For example, group 1 completed both anchor texts, short text 1, long texts 11 and 20 with the accompanying MC test items. Group 2 overlapped with group 1 for the anchor texts, short text 1 and long text 20. Group 2 differed from group 1 in completing long text 2 instead of long text 11.

items to new items (i.e., 32 to 40% anchor items in relation to 68 to 60% new items) is in line with the recommendation to include at least 20% anchor items (Kolen and Brennan, 2014). These new texts (with accompanying MC items) were carefully selected to compile multiple test forms that are as equal as possible (e.g., including both short and long texts, different topics). To ensure that each text would be read by a minimum of 300 students, an overlap in texts across the 20 groups was provided. Moreover, the order in which the texts was presented was counterbalanced to reduce the influence of participants' fatigue and motivation.

4.1.5. Step 5: Pilot study

In a fifth step, a pilot study was conducted in one class of fifth graders ($n = 22$) to examine (a) the required time for completing the test, (b) the comprehensibility, and (c) an initial exploration of the difficulty level of the items (i.e., see the section 'psychometric validation' for an extensive evaluation of the item difficulty based on the subsequent large-scale study). These participants completed the 21 MC items belonging to the anchor texts. The anchor items and texts were representative for the other included texts and items (e.g., comparable text difficulty and length, variety in comprehension levels of the items). Different test items were adapted, reformulated, or replaced by new items based on students' feedback.

4.1.6. Step 6: Large-scale study

4.1.6.1. Test administration and test scoring

During two assessment occasions, the paper and pencil version of the reading comprehension test was administered to a large student sample (see participant section). During the first test moment, the purpose of the study was explained and both anchor texts with 21 MC items were completed by all students. During the second test moment, the three or four additional texts with the accompanying MC items different per group of 150 students were administered. Based on practical

considerations, students from the same class were assigned to the same group. However, within the same school, the allocation to different groups was varied to limit school effects. In addition, classes from fifth and sixth grade were distributed at random among the 20 groups.

Before the start of each assessment occasion, the following instructions were given: (a) the test consists of multiple text passages and accompanying MC items; (b) only one out of the three answer options is correct; (c) anything, except a dictionary, you normally use during reading a text may but must not be used (e.g., highlighter, scratch paper); and (d) time limitation (i.e., 20 min for both anchor texts and 30 min for the additional texts varying per group). A time limitation was included, since the intention of reading comprehension is not only achieving a high comprehension level but also regulating the own comprehension as efficiently as possible (Leopold and Leutner, 2012). Instructions were described in detail in the protocol to standardize the administration process. Both the main researcher and trained research assistants followed this standardized protocol carefully to administrate the reading comprehension test. Further, standard answers on possible questions were also included in the protocol ordered by don'ts (e.g., explaining words of the reading passage, using a dictionary) and do's (e.g., explaining words in the MC items like 'paragraph' or 'reliable information').

As to the test scoring, the MC items were dichotomously scored (0 = incorrect; 1 = correct), with one correct answer per test item.

4.1.6.2. Participants

In total 3,269 late elementary Flemish (Belgium) students from 167 classes from 68 different schools participated in the study. Participants were recruited via convenience sampling. The mean age of the students was 11.32 years (SD = 0.68), with a minimum of 9.0 and a maximum of 14.9. Of the students, 51.6% were fifth graders and 48.4% were sixth graders. Further, 49.4% of the students were girls and 50.6% were boys. 5.0% of the students were diagnosed with dyslexia. 87.2% of the students were indicated by the teachers as Dutch-speaking students, 7.2% as non-native students, and 5.6% as bilingual students (i.e., speaking Dutch and another language at home). Informed consent was obtained for all participants (i.e., students, parents, teachers, and school principals), in line with the country's privacy legislation.

4.1.6.3. Psychometric validation

The psychometric specifications consist of the desired statistical properties of the items (e.g., item difficulty and item discrimination) and these of the whole test (e.g., test reliability and test difficulty; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education et al., 2014). Both classical test theory (CTT) and item response theory (IRT), considered as complementary psychometric approaches (De Champlain, 2010), were conducted to evaluate the psychometric quality of both the test items and the complete RC-PM tool. In this respect, CTT - a test-level theory - is mainly useful in the early phases of processing, while IRT - an item-level theory - can be applied to estimate final statistical properties of test items (De Champlain, 2010; Demars, 2018). CTT and IRT were conducted in R 3.6.1., respectively by means of the TAM (Robitzsch et al., 2018) and mirt (Chalmers, 2012) package. IRT has already been applied for many large language tests (Davidson, 2004). For example, in the study of Bourdeaud'Hui et al. (2021) IRT was found to be valuable to validate the psychometric quality of a comprehensive listening test. Besides test and item validation, automated test assembly (ATA) was conducted in R 3.6.1. (rata-package) to create parallel tests forms that are equivalent in difficulty and length (Li et al., 2021). Below, the different psychometric steps that were taken are described in detail. Due to the large number of test forms and the relatively small overlap between them, an IRT scale was first calibrated using the items all students completed (i.e., the anchor items). Afterwards, the items of the remaining 33 texts were put on the same scale.

First, CTT was conducted to evaluate the item parameters, i.e., item difficulty and item discrimination for each of the anchor items. Item difficulty is calculated as the proportion of students (p-value) that has successfully completed an item. p-values of 1.00 indicate very easy items that are answered correctly by all students, whereas p-values of 0.00 indicate very difficult items that are answered incorrectly by all students. Item discrimination is calculated as Pearson item-total correlation, as such it refers to the degree to which the performance on an item correlates with the performance on the total test. Items with a negative or item-total correlation value below 0.20 were removed from the analysis (Varma, 2006).

Second, the degree to which the data of the anchor items fit a 2-parameter logistic model was investigated. Further, item misfit was investigated using the S_X2 statistics (Orlando and Thissen, 2003). As to the model fit, values of the RMSEA (i.e., root mean square error of approximation) and SRMSR (i.e., standardized root mean square residuals) should lie below 0.06 (Hu and Bentler, 1999) and values higher than 0.90 for CFI (i.e., comparative fit indices), and TLI (Tucker-Lewis index; Little, 2013).

Third, the items of all remaining texts and the anchor texts were put on the same ability scale. For this purpose, the items of each remaining text were separately linked to the anchor items. Once again, CTT was used to estimate the item parameters of the items of each remaining text, combined with the two anchor texts. Further, the 2PL model was used to estimate the item parameters of each remaining text. In order to put all items of the remaining texts on the same ability scale, the item parameters of the anchor items were constrained and used to estimate the parameters of the items of each remaining text separately. Item fit statistics were estimated to select all items of each remaining text fitting with the anchor items and model fit was estimated for each new combination of items.

Fourth, once all selected items were put on the same ability scale, automated test assembly was conducted to construct six test forms that are equivalent in test difficulty and length (Diao and van der Linden, 2011). For each test form, maximum test information (i.e., test reliability) within the ability scale was also intended.

Finally, the reliability of the six test forms was calculated. In this respect, a reliability coefficient of 0.50 or 0.60 is considered as sufficient for classrooms tests (Rudner and Schafer, 2002). More specifically, coefficients below 0.50 refer to low reliability, 0.50 to 0.80 to moderate reliability and above 0.80 to high reliability.

5. Results

5.1. Classical test theory analysis: Anchor items

Regarding the item discrimination of the anchor items, the results in Table 4 indicate that the item-total correlation of 2 of the 21

items was located below 0.20. As both these items do not discriminate enough between students, they were not retained for further analysis. Regarding the item difficulty of the anchor items, no *p*-values were too close to zero or one (Table 4). Since the anchor texts did not contain too difficult or too easy items, no additional items were removed.

5.2. Item response theory analysis: Anchor items

Regarding item misfit, the *p*-values of the S_X^2 statistics in Table 5 show that the observed data of 14 out of the 19 remaining anchor items do not differ significantly from the responses predicted by the model. As such, these items were retained for further analysis. Further, Table 5 includes item discrimination and difficulty for each item. Considering model fit, the values of the RMSEA and SRMSR lie below 0.06 and the values of the TLI and CFI are higher than 0.90, indicating that a unidimensional ability scale with 14 of the 21 anchor items fit the data well ($\chi^2 = 183.73$, $df = 77$, $p < 0.001$, RMSEA = 0.02, SRMSR = 0.02, TLI = 0.94, CFI = 0.95).

5.3. Item and model fit for the items of the remaining 33 texts

In a next phase of the analysis, we evaluated the degree to which the items of each remaining text could be calibrated on the same ability scale of the anchor items.

First, CTT was conducted to estimate the item parameters of each remaining text compared with the anchor texts at the level of the whole test. Next, based on IRT analyses, all other items of the

TABLE 4 Item-total correlations of the anchor items.

Item	Item-total correlation	*p*-value	Item	Item-total correlation	*p*-value
1	0.406	0.748	12	0.288	0.856
2	0.434	0.771	13	0.277	0.494
3	0.450	0.685	14	0.414	0.779
4	0.386	0.526	15	0.326	0.495
5	0.375	0.611	16	0.377	0.719
6	**0.157**	0.243	17	0.219	0.807
7	0.375	0.761	18	0.356	0.645
8	0.417	0.734	19	0.345	0.432
9	0.328	0.922	20	0.351	0.746
10	0.344	0.623	21	**0.138**	0.483
11	0.278	0.473			

The bold values indicate that the item-total correlation of these values is below 0.20.

TABLE 5 Fit indices of the anchor items.

Item	Difficulty	Discrimination	S_X^2	Df S_X^2	RMSEA S_X^2	*p* S_X^2
1	1.31	1.01	15.34	9	0.02	0.45
2	1.55	1.22	8.10	9	0.00	0.64
3	0.90	0.88	12.90	9	0.01	0.45
4	0.50	0.67	12.55	9	0.01	0.45
5	3.31	1.53	11.28	9	0.01	0.45
6	0.54	0.55	5.55	10	0.00	0.85
7	−0.11	0.33	10.19	9	0.01	0.47
8	1.96	0.74	11.63	10	0.01	0.47
9	−0.03	0.28	8.23	10	0.00	0.65
10	1.06	0.79	14.46	10	0.01	0.45
11	1.49	0.44	12.89	10	0.01	0.45
12	0.66	0.69	7.83	9	0.00	0.64
13	−0.29	0.48	16.53	9	0.02	0.45
14	1.24	0.86	13.58	10	0.01	0.45

remaining 33 texts were scattered on the same ability scale, compared with the fixed anchor items of the previous step. In this respect, the fit indices of these items together with the anchor items were investigated. Items with an item-total correlation below 0.20 were removed. Items that differed significantly from the responses predicted by the model were removed as well (i.e., items with a significant p-value of the S_X^2 statistics). As to the fit indices of the anchor model, values below 0.06 for RMSEA and SRMSR and values higher than 0.90 for TLI and CFI were used as benchmark in assessing the fit indices for the model of the remaining 33 texts with their accompanying MC items compared with the anchor model. In total, 270 items of the initial 363 items remained (see Appendix).

5.4. Test assembling

The 270 remaining items of the calibrated item pool were combined into six parallel test forms (test assembly) that are equivalent in difficulty and length (Diao and van der Linden, 2011). As a first test specification, maximization of the test information function for a medium ability level was entered as the objective function. To minimize the administration time of each test form, a test length of 15 to 20 items was specified as constraint object. Further, constraint specifications were set to guarantee (1) no overlapping items between test forms; (2) that items of a specific text were not distributed across multiple test forms. A mixed-integer programming (MIP)-solver was used to find an optimal solution for the given combination of constraints and the objective. Table 6 shows the remaining 107 MC items and their parameters under the 2PLM, distributed across six test forms. Each of these tests consists of 17 to 18 MC items belonging to two or three texts. Each text was covered in only one of the six test forms. Further, the MC items of each test form covered more or less evenly the three comprehension levels. Table 7 presents an overview of the six test forms. Figure 2 presents the test information functions of the six equivalent test forms at a given ability level θ. Figure 3 presents the test characteristics curve of the six equivalent test forms.

5.5. Reliability

The reliability of the six test forms is conditional, varying across different points on the ability scale. For an ability level of −0.6, the reliability for the six test forms is around 0.66. For an ability level of 1.8, the reliability of the six test forms is between 0.66 and 0.77. At the moderate ability level, the reliability ranges from 0.77 to 0.80. These values indicate a moderate to good internal consistency of each of the six test forms.

6. Discussion and conclusion

Although reading comprehension is considered a key competence in today's society, it appears to be challenging for many late elementary students (Rasinski, 2017). Simultaneously, the end of elementary school in transition to secondary education appears however a critical period in the development of this skill (Keresteš et al., 2019). Monitoring students' reading comprehension in order to make effective and tailored instructional decisions is therefore required (Stecker et al., 2008; Zeuch et al., 2017). However, appropriate reading comprehension monitoring instruments are lacking. The present study therefore responds to the call for theoretical and empirical substantiated comprehension monitoring instruments by developing the Reading Comprehension – Progress Monitoring tool (RC-PM tool) for late elementary students.

The undertaken developed process and final composition of the RC-PM tool entails various strengths which can be related to the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). First, content validity was guaranteed through a comprehensive construct analysis, alignment with the Flemish reading comprehension curriculum, structured development of test items based on a template, an expert panel review, and a pilot study. Second, internal validity of the RC-PM tool was examined through a psychometric validation process. More specifically, both classical test theory and item response theory analyses were conducted. Automated test assembly was applied to compile six parallel test forms equivalent in difficulty and length. A moderate to good internal consistency of these six tests forms was found. In this respect, this is the first study so far composing a test comprising six equivalent theoretically and empirically grounded parallel test forms for assessing reading comprehension of expository texts on regular basis in late elementary education.

Notwithstanding this study's relevant contribution, researchers are advised to explore the inclusion of multiple assessments formats in further research, such as combining MC-items with open-ended items (Campbell, 2005), think aloud tasks (McMaster et al., 2012, 2014), or observations. As to the latter, it is remarkable that – to our knowledge – observations are mostly used to map teachers' instruction and not students' reading comprehension (e.g., Brevik, 2019; Magnusson et al., 2019). Due to its scope, this was not feasible in the current research (e.g., observe or evaluate open-ended questions of a large group of participants). However, using multiple assessment formats to map reading comprehension progress in the future can make the process of identifying especially struggling comprehenders more accurate (Munger and Blachman, 2013). In this respect, think aloud tasks, mapping the comprehension process instead of the product, can be used to unravel the specific elements with which students struggle (McMaster et al., 2012, 2014). An additional suggestion for future research is to explore the external validity of the RC-PM tool. More specifically, it is recommended to compare students' scores on the RC-PM tool with their performance mapped via other reading comprehension tests (e.g., the Cito test for reading comprehension; the emerging Flemish central tests). Further, it is recommended in future research to examine the effect of offering a time limitation, since test speediness could be a cofounding factor. A next limitation is related to the psychometric validation of the RC-PM tool. In this regard, the dimensionality of each of the items was not considered, which would be valuable in future research. More specifically, it would be recommended to evaluate how the items of the parallel test forms meet the model assumptions (i.e., unidimensionality, local independence, monotonicity) if the

10.3389/feduc.2023.1066837

TABLE 6 Overview of the six test forms.

Test form	Item	Subskill	Text type	Discrimination	Difficulty
1	Text5 _Item4	Surface model (meaning at micro level)	Long expository text	0,58	−0,19
1	Text5 _Item6	Surface model (replacement)	Long expository text	0,82	0,38
1	Text5 _Item8	Textbase model (text structure)	Long expository text	0,52	−0,55
1	Text5 _Item9	Textbase model (referrals)	Long expository text	1,3	2,54
1	Text5_Item1	Surface model (meaning at micro level)	Long expository text	1,11	0,39
1	Text5_Item10	Textbase model (summerazing)	Long expository text	0,22	−0,17
1	Text5_Item11	Situation model (evaluating reliability)	Long expository text	1,37	0,41
1	Text5_Item12	Situation model (take a different view)	Long expository text	0,21	−0,94
1	Text5_Item2	Situation model (text purpose)	Long expository text	0,79	−0,09
1	Text5_Item3	Situation model (predicting)	Long expository text	0,41	0,39
1	Text8_Item1	Textbase model (referrals)	Short expository text	0,49	0,19
1	Text8_Item2	Textbase model (meaning at macro level)	Short expository text	0,28	1,09
1	Text8_Item3	Situation model (text purpose)	Short expository text	0,7	0,1
1	Text8_Item4	Surface model (expression)	Short expository text	0,63	1,17
1	Text8_Item5	Textbase model (summerazing)	Short expository text	0,43	1,87
1	Text8_Item6	Textbase model (text structure)	Short expository text	0,08	−0,66
1	Text8_Item7	Situation model (evaluating reliability)	Short expository text	1,33	3,03
1	Text8_Item8	Surface model (meaning at micro level)	Short expository text	0,41	1,76
2	Text7_Item10	Textbase model (text structure)	Long expository text	1,49	2
2	Text7_Item11	Surface model (replacement)	Long expository text	0,84	1,38
2	Text7_Item4	Textbase model (meaning at macro level)	Long expository text	0,86	0,04
2	Text7_Item6	Surface model (meaning at micro level)	Long expository text	0,55	−0,16
2	Text7_Item8	Situation model (evaluating reliability)	Long expository text	0,77	0,4
2	Text7_Item9	Situation model (text purpose)	Long expository text	0,67	−0,12
2	Text31_Item1	Textbase model (signal words)	Long expository text	0,69	0,08
2	Text31_Item10	Textbase model (refferals)	Long expository text	0,52	−0,64
2	Text31_Item11	Surface model (replacement)	Long expository text	0,23	−0,26

(Continued)

10.3389/feduc.2023.1066837

TABLE 6 (Continued)

Test form	Item	Subskill	Text type	Discrimination	Difficulty
2	Text31_Item12	Textbase model (meaning at macro level)	Long expository text	0,16	−0,33
2	Text31_Item2	Surface model (meaning at micro level)	Long expository text	1,26	1,05
2	Text31_Item3	Surface model (replacement)	Long expository text	0,93	2,01
2	Text31_Item4	Situation model (take a different view)	Long expository text	0,57	0,03
2	Text31_Item5	Situation model (evaluating reliability)	Long expository text	0,15	−0,65
2	Text31_Item6	Textbase model (summerazing)	Long expository text	0,83	−1,19
2	Text31_Item7	Textbase model (text structure)	Long expository text	−0,08	−1,32
2	Text31_Item8	Situation model (text purpose)	Long expository text	0,38	0,18
2	Text31_Item9	Situation model (predicting)	Long expository text	0,67	0,54
3	Anchor1_Item1	Surface model (meaning at micro level)	Short expository text	1,01	1,31
3	Anchor1_Item2	Textbase model (meaning at macro level)	Short expository text	1,22	1,55
3	Anchor1_Item3	Textbase model (referrals)	Short expository text	0,88	0,9
3	Anchor1_Item5	Surface model (replacement)	Short expository text	0,67	0,5
3	Text1_Item3	Surface model (replacement)	Short expository text	0,43	−1,22
3	Text1_Item5	Surface model (replacement)	Short expository text	0,79	1,09
3	Text1_Item7	Situation model (predicting)	Short expository text	0,78	0,96
3	Text30_Item1	Surface model (expression)	Long expository text	0,84	0,67
3	Text30_Item10	Surface model (replacement)	Long expository text	0,52	−0,35
3	Text30_Item11	Situation model (evaluating reliability)	Long expository text	0,43	−0,19
3	Text30_Item12	Textbase model (referrals)	Long expository text	0,18	−0,4
3	Text30_Item3	Textbase model (text structure)	Long expository text	0,32	−0,04
3	Text30_Item4	Surface model (expression)	Long expository text	0,45	−0,71
3	Text30_Item5	Textbase model (meaning at macro level)	Long expository text	0,43	0,21
3	Text30_Item6	Situation model (text purpose)	Long expository text	0,71	−0,35
3	Text30_Item7	Situation model (predicting)	Long expository text	0,97	0,25
3	Text30_Item8	Situation model (take a different view)	Long expository text	0,24	0,27

(Continued)

TABLE 6 (Continued)

Test form	Item	Subskill	Text type	Discrimination	Difficulty
4	Text17_Item10	Surface model (replacement)	Long expository text	0,85	1,95
4	Text17_Item11	Textbase model (text structure)	Long expository text	0,85	0,08
4	Text17_Item12	Surface model (expression)	Long expository text	0,85	0,35
4	Text17_Item2	Situation model (evaluating reliability)	Long expository text	0,85	1,47
4	Text17_Item5	Situation model (text purpose)	Long expository text	0,85	0,74
4	Text17_Item7	Situation model (take a different view)	Long expository text	0,85	1,36
4	Text22_Item1	Surface model (replacement)	Short expository text	0,52	0,42
4	Text22_Item2	Textbase model (text structure)	Short expository text	0,45	0
4	Text22_Item3	Textbase model (referrals)	Short expository text	0,27	0,24
4	Text22_Item4	Surface model (replacement)	Short expository text	0,69	−0,06
4	Text22_Item5	Textbase model (summerazing)	Short expository text	0,25	−0,1
4	Text22_Item7	Situation model (predicting)	Short expository text	0,47	0,8
4	Text22_Item8	Situation model (text purpose)	Short expository text	1,44	3,59
4	Text28_Item2	Surface model (replacement)	Short expository text	0,17	0,93
4	Text28_Item4	Surface model (meaning at micro level)	Short expository text	0,62	−0,11
4	Text28_Item5	Textbase model (summerazing)	Short expository text	0,18	−0,1
4	Text28_Item6	Situation model (text purpose)	Short expository text	0,88	2,86
4	Text28_Item7	Situation model (take a different view)	Short expository text	1,01	0,01
5	Text14_Item10	Surface model (replacement)	Long expository text	0,25	−0,22
5	Text14_Item11	Textbase model (text structure)	Long expository text	0,56	0,33
5	Text14_Item12	Textbase model (referrals)	Long expository text	0,68	−0,57
5	Text14_Item2	Surface model (meaning at micro level)	Long expository text	0,2	−1,18
5	Text14_Item4	Surface model (expression)	Long expository text	0,44	−0,47
5	Text14_Item5	Situation model (text purpose)	Long expository text	0,52	0,59
5	Text14_Item8	Situation model (predicting)	Long expository text	0,28	−1,11
5	Text14_Item9	Situation model (take a different view)	Long expository text	1,25	1,03

(Continued)

10.3389/feduc.2023.1066837

TABLE 6 (Continued)

Test form	Item	Subskill	Text type	Discrimination	Difficulty
5	Text32_Item10	Textbase model (summerazing)	Long expository text	0,52	0,79
5	Text32_Item11	Surface model (replacement)	Long expository text	0,4	0,21
5	Text32_Item12	Textbase model (summerazing)	Long expository text	0,76	−0,16
5	Text32_Item2	Surface model (meaning at micro level)	Long expository text	1,44	1,56
5	Text32_Item3	Textbase model (referrals)	Long expository text	0,31	−0,29
5	Text32_Item4	Textbase model (text structure)	Long expository text	0,69	−0,88
5	Text32_Item5	Situation model (take a different view)	Long expository text	0,99	0,53
5	Text32_Item7	Situation model (text purpose)	Long expository text	0,81	0,59
5	Text32_Item8	Situation model (take a differnt view)	Long expository text	0,44	0,56
5	Text32_Item9	Situation model (text purpose)	Long expository text	0,19	−0,25
6	Text18_Item1	Surface model (replacement)	Long expository text	0,29	−1,39
6	Text18_Item12	Textbase model (referrals)	Long expository text	0,2	−0,68
6	Text18_Item3	Textbase model (referrals)	Long expository text	0,31	−0,2
6	Text18_Item4	Surface model (meaning at micro level)	Long expository text	0,57	0,54
6	Text18_Item6	Surface model (expression)	Long expository text	0,95	−0,05
6	Text18_Item7	Textbase model (summerazing)	Long expository text	1,23	0,77
6	Text18_Item9	Situation model (predicting)	Long expository text	0,43	−0,09
6	Text20_Item1	Textbase model (signal words)	Long expository text	0,85	1,31
6	Text20_Item10	Surface model (meaning at mciro level)	Long expository text	0,49	0,73
6	Text20_Item11	Surface model (replacement)	Long expository text	1,1	1,66
6	Text20_Item12	Surface model (replacement)	Long expository text	0,6	−1,18
6	Text20_Item3	Textbase model (meaning at macro level)	Long expository text	0,37	−1,5
6	Text20_Item4	Textbase model (meaning at macro level)	Long expository text	0,68	−0,39
6	Text20_Item5	Textbase model (summerazing)	Long expository text	0,5	−0,42
6	Text20_Item6	Situation model (predicting)	Long expository text	0,4	0,65
6	Text20_Item7	Situation model (text purpose)	Long expository text	0,22	−0,93
6	Text20_Item8	Situation model (evaluating reliability)	Long expository text	0,98	0,51
6	Text20_Item9	Situation model (predicting)	Long expository text	0,54	0,39

TABLE 7 Overview of the six test forms.

Test form	Texts		Items			
	Number of texts	Total word count	Surface	Textbase	Situation	Total
1	2	393	5	7	6	18
2	2	550	5	7	6	18
3	3	519	7	5	5	17
4	3	516	6	5	7	18
5	2	468	5	6	7	18
6	2	526	6	7	5	18

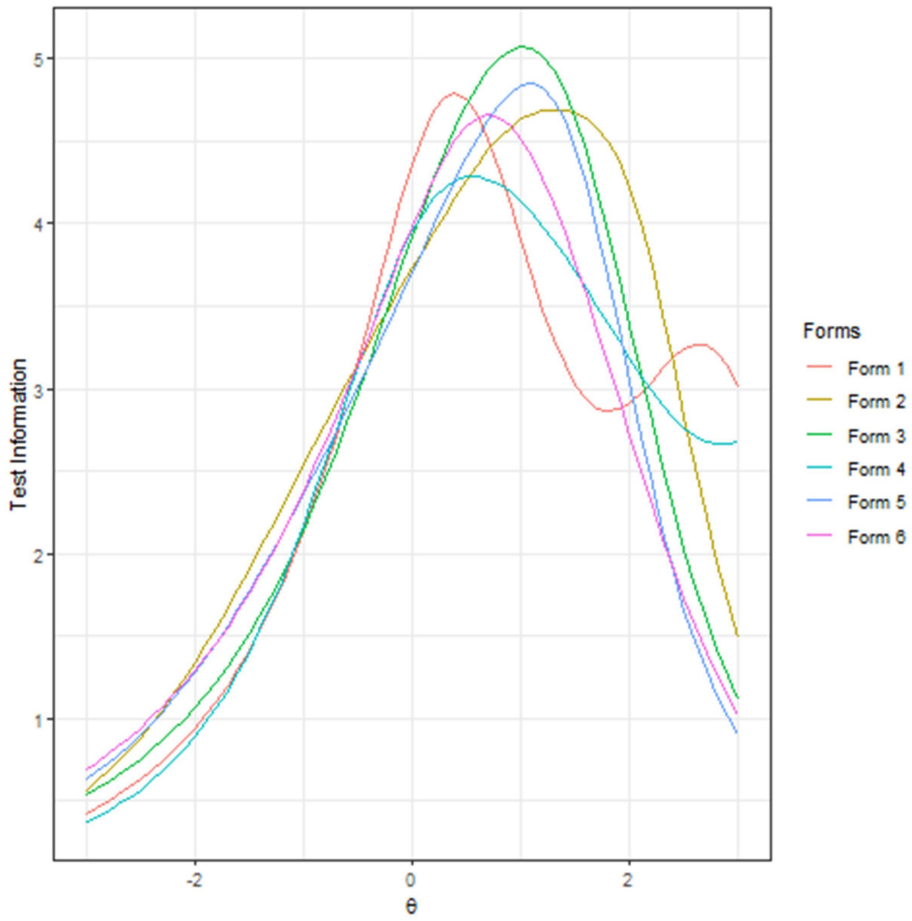

FIGURE 2
Test information functions at a given ability level θ of six equivalent test forms.

parallel test forms will be used in future research to estimate students' abilities. Finally, the RC-PM tool was developed in the specific context of Flanders. It would be interesting to evaluate the usability of this (translated) tool beyond the specific context of this study.

Next, to the advice for future research described above, we conclude with some important implications of the current study for research and practice. First, the RC-PM tool responds to the lack of appropriate evidence-based progress monitoring tools to map late elementary students' reading comprehension development on a regular basis. For educational test developers, this study provides usability in detail in the steps to be taken to develop a progress monitoring tool in a theoretically and empirically grounded manner. Further, the RC-PM tool can be applied in future research as a valid progress monitoring assessment tool. For example, the RC-PM tool can be applied within the context of response-to-intervention research (RTI), an emerging educational framework within the context of our increasingly diverse society (Jimerson

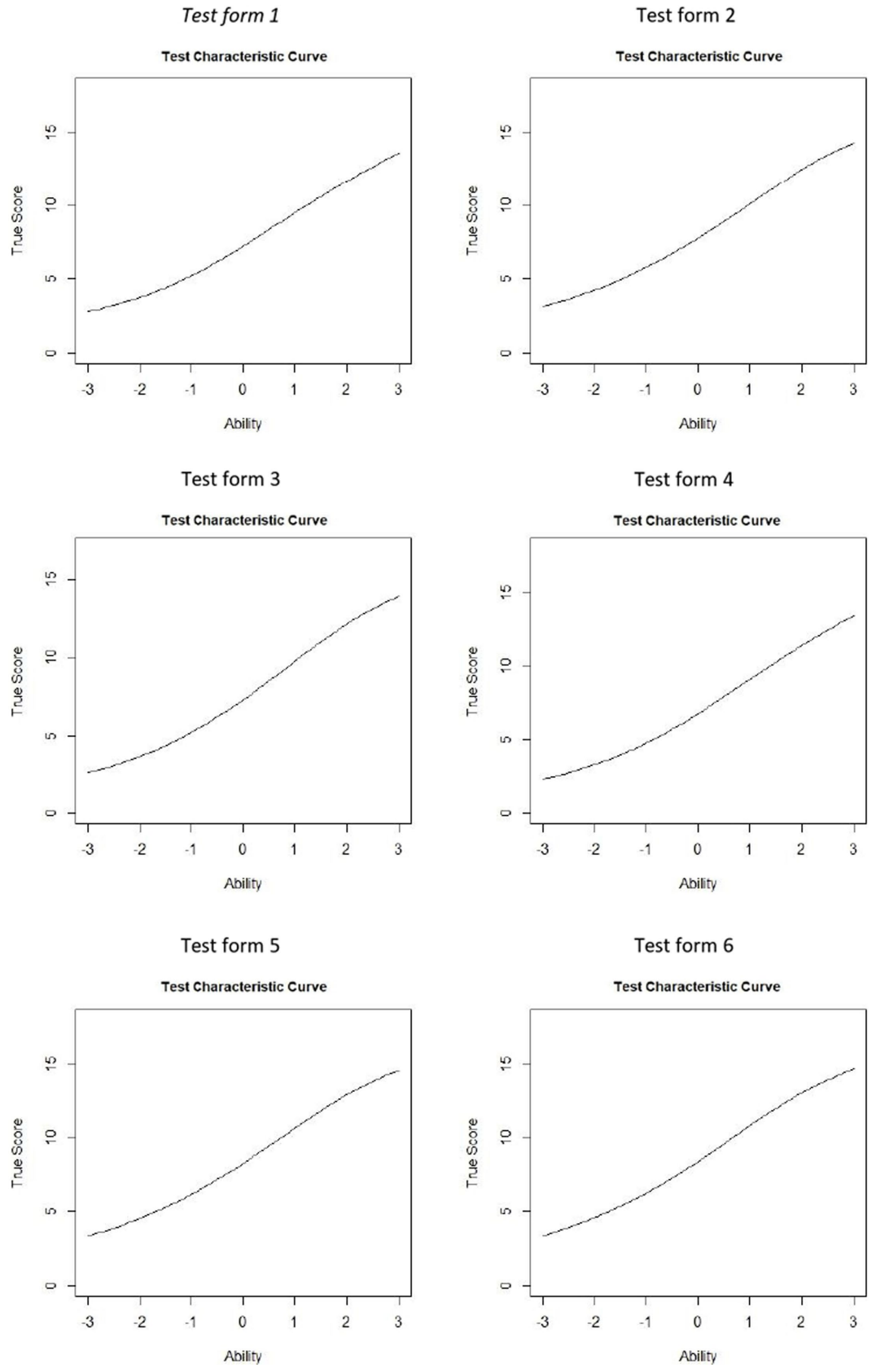

FIGURE 3
Test characteristic curve of six equivalent test forms.

et al., 2016; Jefferson et al., 2017). RTI refers to a multi-tiered approach to identify and support struggling students through purposefully providing targeted varying levels of support (i.e., whole-class instruction, small-group instruction, or individualized instruction; Kaminski and Powell-Smith, 2017). Remarkably, current RTI research focuses more on reading fluency (e.g., Griffiths et al., 2009; Svensson et al., 2019) or uses the criticized cloze task (e.g., Vaughn and Fletcher, 2012; Roberts et al., 2013), possibly due to the lack of appropriate assessment instruments. Our study provides an answer to this concern. Since progress monitoring assessments play a crucial role in the effective implementation of RTI research, the RC-PM tool can be useful within the RTI context (Tolar et al., 2014). As to practice, the short administration time (i.e., 20 to 30 min), the easy test scoring (i.e., one correct answer option per question), and the incorporation of and mapping different comprehension levels in the RC-PM tool, make this tool user-friendly and valuable to implement in classroom practice. The results derived from the RC-PM tool provide insight into the progress students are making. Furthermore, this can serve as starting point for practitioners to align and tailer their instruction to students' learning needs and in this way to integrate assessment, teaching, and learning.

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

RB, EM, and HK designed the study. RB was in charge of the data collection and wrote the main part of the manuscript. KA conducted the data analysis. All authors contributed throughout the different writing stages of the manuscript 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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1066837/full#supplementary-material

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.

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. School Psychol.* 51, 1–18. doi: 10.1016/j.jsp.2012.09.004

Barbe, B. W. (1958). Measuring reading comprehension. *Clearing House* 32, 343–345. doi: 10.1080/00098655.1958.11476143

Bourdeaud'Hui, H., Aesaert, K., and van Braak, J. (2021). Exploring the validity of a comprehensive listening test to identify differences in primary school students' listening skills. *Lang. Assess. Quar.* 18, 228–252. doi: 10.1080/15434303.2020.1860059

Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Educ. Pract.* 5, 238–242. doi: 10.1016/j.nepr.2004.12.005

Brevik, L. M. (2019). Explicit reading strategy instruction or daily use of strategies? Studying the teaching of reading comprehension through naturalistic classroom observation in English L2. *Read. Writing* 32, 2281–2310. doi: 10.1007/s11145-019-09951-w

Cain, K., Oakhill, J., and Bryant, P. (2004). Children's reading comprehension ability: concurrent prediction by working memory, verbal ability, and component skills. *J. Educ. Psychol.* 96, 31–42. doi: 10.1037/0022-0663.96.1.31

Calet, N., López-Reyes, R., and Jiménez-Fernández, G. (2020). Do reading comprehension assessment tests result in the same reading profile? A study of Spanish primary school children. *J. Res. Read.* 43, 98–115. doi: 10.1111/1467-9817.12292

Campbell, J. R. (2005). "Single instrument, multiple measures: considering the use of multiple item formats to assess reading comprehension" in *Children's Reading Comprehension and Assessment*. eds. S. G. Paris and S. A. Stahl (Mahwah, NJ: Lawrence Erlbaum Associates, Inc), 347–368.

Castles, A., Rastle, K., and Nation, K. (2018). Ending the reading wars: reading acquisition from novice to expert. *Psychol. Sci. Public Int.* 19, 5–51. doi: 10.1177/1529100618772271

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48, 1–29. doi: 10.18637/jss.v048.i06

Cito, B.V. (2014). Begrijpend lezen 3.0 voor groep 4 [Reading Comprehension 3.0 for second grade]. Arnhem, the Netherlands: Cito B.V.

Collins, A. A., Lindström, E. R., and Compton, D. L. (2018). Comparing students with and without reading difficulties on reading comprehension assessments: a meta-analysis. *J. Learn. Disabil.* 51, 108–123. doi: 10.1177/0022219417704636

Davidson, F. (2004). The identity of language testing. *Lang. Assess. Q.* 1, 85–88. doi: 10.1207/s15434311laq0101_9

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika* 9, 185–197. doi: 10.1007/BF02288722

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Med. Educ.* 44, 109–117. doi: 10.1111/j.1365-2923.2009.03425.x

Demars, C. E. (2018). "Classical test theory and item response theory" in *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*. eds. P. Irwing, T. Booth and D. J. Hughes, vol. *2* (Hoboken, NJ: John Wiley & Sons Ltd.), 49–73.

Deno, S. L. (1985). Curriculum-based measurement: the emerging alternative. *Except. Children* 52, 219–232. doi: 10.1177/001440298505200303

Diao, Q., and van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in r. *Appl. Psychol. Meas.* 35, 398–409. doi: 10.1177/0146621610392211

Dorans, N. J., Moses, T. P., and Eignor, D. R. (2010). *Principles and Practices of Test Score Equating (ETS Research Report No. RR-10-29)*. Princeton, NJ: Educational Testing Service.

Elleman, A. M., and Oslund, E. L. (2019). Reading comprehension research: Implications for practice and policy. *Policy Insights Behav. Brain Sci.* 6, 3–11. doi: 10.1177/2372732218816339

Follmer, D. J., and Sperling, R. A. (2018). Interactions between reader and text: Contributions of cognitive processes, strategy use, and text cohesion to comprehension of expository science text. *Learn. Individ. Diff.* 67, 177–187. doi: 10.1016/j.lindif.2018.08.005

Förster, N., Kawohl, E., and Souvignier, E. (2018). Short- and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learn. Instruct.* 56, 98–109. doi: 10.1016/j.learninstruc.2018.04.009

Förster, N., and Souvignier, E. (2011). Curriculum-based measurement: developing a computer-based assessment instrument for monitoring student reading progress on multiple indicators. *Learning Disabil. Comtemp. J.* 9, 21–44.

Förster, N., and Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learn. Instruct.* 32, 91–100. doi: 10.1016/j.learninstruc.2014.02.002

Förster, N., and Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychol. Rev.* 44, 60–75. doi: 10.17105/SPR44-1.60-75

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

Goegan, L. D., Radil, A. I., and Daniels, L. M. (2018). Accessibility in questionnaire research: integrating universal design to increase the participation of individuals with learning disabilities. *Learn. Disabil. Contemp. J.* 16, 177–190. doi: 10.7939/r3-cmkq-1c82

Green, R. (2017). *Designing Listening Tests: A Practical Approach*. London: Palgrave Macmillan.

Griffiths, A. J., VanDerHeyden, A. M., Skokut, M., and Lilles, E. (2009). Progress monitoring in oral reading fluency within the context of RTI. *School Psychol. Q.* 24, 13–23. doi: 10.1037/a0015435

Hacquebord, H., Stellingwerf, B., Linthorst, R., and Andringa, S. (2005). *Diataal: Verantwoording En Normering [Diataal: Manual]*. Groningen, The Netherlands: Rijksuniversiteit Groningen.

Haladyna, T. M., and Rodriguez, M. C. (2013). *Developing and Validating Test Items*. Milton Park: Routledge.

Harlaar, N., Dale, P. S., and Plomin, R. (2007). From learning to read to reading to learn: Substantial and stable genetic influence. *Child Dev.* 78, 116–131. doi: 10.1111/j.1467-8624.2007.00988.x

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., et al. (2021). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Lang. Testing* 38, 41–61. doi: 10.1177/0265532220917316

Hoover, H. D., Dunbar, S. B., and Frisbie, D. A. (2003). *The Iowa Tests: Guide to Research and Development*. Itasca, IL: Riverside.

Houtveen, A. A. M., van Steensel, R. C. M., and de la Rie, S. (2019). De vele kanten van leesbegrip. Literatuurstudie naar onderwijs in begrijpend lezen in opdracht van het Nationaal Regieorgaan Onderwijsonderzoek en de Inspectie van het Onderwijs. Available at: http://hdl.handle.net/1765/119125 (Accessed September 8, 2020).

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equat. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Jefferson, R. E., Grant, C. E., and Sander, J. B. (2017). Effects of Tier I differentiation and reading intervention on reading fluency, comprehension, and high stakes measures. *Read. Psychol.* 38, 97–124. doi: 10.1080/02702711.2016.1235648

Jensen, K. L., and Elbro, C. (2022). Clozing in on reading comprehension: a deep cloze test of global inference making. *Reading Writing* 35, 1221–1237. doi: 10.1007/s11145-021-10230-w

Jimerson, S. R., Burns, M. K., and Van Der Heyden, A. M. (2016). *Handbook of Response to Intervention 2nd Edn*, New York: Springer.

Kaminski, R. A., and Powell-Smith, K. A. (2017). Early literacy intervention for preschoolers who need Tier 3 support. *Topics Early Childhood Spec. Educ.* 36, 205–217. doi: 10.1177/0271121416642454

Keenan, J. M., Betjemann, R. S., and Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Sci. Stud. Read.* 12, 281–300. doi: 10.1080/10888430802132279

Kendeou, P., Van Den Broek, P., Helder, A., and Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learn. Disabil. Res. Pract.* 29, 10–16. doi: 10.1111/ldrp.12025

Keresteš, G., Brkovic, I., Siegel, L. S., Tjus, T., and Hjelmquist, E. (2019). Literacy development beyond early schooling: a 4-year follow-up study of Croatian. *Reading Writing* 32, 1955–1988. doi: 10.1007/s11145-018-9931-9

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: The CI perspective. *Discourse Processes* 39, 125–128. doi: 10.1207/s15326950dp3902&3_2

Kolen, J. M., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking. Methods and Practices. 3RD Edn*, New York: Springer.

Lee, K., and Chen, X. (2019). An emergent interaction between reading fluency and vocabulary in the prediction of reading comprehension among French immersion elementary students. *Reading Writing* 32, 1657–1679. doi: 10.1007/s11145-018-9920-z

Leopold, C., and Leutner, D. (2012). Science text comprehension: Drawing, main idea selection, and summarizing as learning strategies. *Learn. Instruct.* 22, 16–26. doi: 10.1016/j.learninstruc.2011.05.005

Leslie, L., and Caldwell, J. (2014). "Formal and Informal Measures of Reading Comprehension" in *Handbook of Research on Reading Comprehension*. eds. S. E. Israel and G. Duffy (Milton Park: Routledge), 427–451.

Li, J. T., Tong, F., Irby, B. J., Lara-Alecio, R., and Rivera, H. (2021). The effects of four instructional strategies on English learners' English reading comprehension: A meta-analysis. *Lang. Teach. Res.* doi: 10.1177/1362168821994133

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York: Guildford Press.

Magnusson, C. G., Roe, A., and Blikstad-Balas, M. (2019). To what extent and how are reading comprehension strategies part of language arts instruction? A study of lower secondary classrooms. *Read. Res. Q.* 54, 187–212. doi: 10.1002/rrq.231

McMaster, K. L., den Broek, P., Espin, C., White, M. J., Rapp, D. N., Kendeou, P., et al. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learn. Individ. Differ.* 22, 100–111. doi: 10.1016/j.lindif.2011.11.017

McMaster, K. L., Espin, C. A., and Van Den Broek, P. (2014). Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learn. Disabil. Res. Pract.* 29, 17–24. doi: 10.1111/ldrp.12026

McNamara, D. S., and Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychol. Learn. Motivat.* 51, 297–384. doi: 10.1016/S0079-7421(09)51009-2

Meneghetti, C., De Beni, R., and Cornoldi, C. (2007). Strategic knowledge and consistency in students with good and poor study skills. *Europ. J. Cogn. Psychol.* 19, 628–649. doi: 10.1080/09541440701325990

Merchie, E., Gobyn, S., De Bruyne, E., De Smedt, F., Schiepers, M., Vanbuel, M., et al. (2019). *Effectieve, eigentijdse begrijpend leesdidactiek in het basisonderwijs. Wetenschappelijk eindrapport van een praktijkgerichte literatuurstudie*. Belgium, Brussel: Vlaamse Onderwijsraad.

Muijselaar, M. M. L., De Bree, E. H., Steenbeek-Planting, E. G., and De Jong, P. F. (2017). Is de cloze-toets een betrouwbare en valide maat voor begrijpend lezen? *Pedagogische Studien* 94, 418–435.

Munger, K. A., and Blachman, B. A. (2013). Taking a "simple view" of the dynamic indicators of basic early literacy skills as a predictor of multiple measures of third-grade reading comprehension. *Psychol. Schools* 50, 722–737. doi: 10.1002/pits.21699

Nundy, S., Kakar, A., and Bhutta, Z. A. (2022). "Developing Learning Objectives and Evaluation: Multiple Choice Questions/Objective Structured Practical Examinations" in *Medicine and Publish from Developing Countries?* (Singapore: Springer), 393–404.

Orlando, M., and Thissen, D. (2003). Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. *Appl. Psychol. Meas.* 27, 289–298. doi: 10.1177/0146621603027004004

Peters, M. T., Hebbecker, K., and Souvignier, E. (2021). Effects of providing teachers with tools for implementing assessment-based differentiated reading instruction in second grade. *Assess. Effect. Interv.* 47, 157–169. doi: 10.1177/15345084211014926

Randi, J., Grigorenko, E. L., and Sternberg, R. J. (2005). "Revisiting definitions of reading comprehension: Just what is reading comprehension anyway?" in *Metacognition in Literacy Learning: Theory, Assessment, Instruction, and Professional Development*. eds. S. E. Israel, C. Collins-Block, K. Bauserman and K. Kinnucan-Welsch (Mahwah: Erlbaum), 19–40.

Rasinski, T. V. (2017). Readers who struggle: Why many struggle and a modest proposal for improving their reading. *Read Teach* 70, 519–524. doi: 10.1002/trtr.1533

Roberts, G., Good, R., and Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychol. Q.* 20, 304–317. doi: 10.1521/scpq.2005.20.3.304

Roberts, G., Vaughn, S., Fletcher, J., Stuebing, K., and Barth, A. (2013). Effects of a response-based, tiered framework for intervening with struggling readers in middle school. *Read. Res. Q.* 48, 237–254. doi: 10.1002/rrq.47

Robitzsch, A., Kiefer, T., and Wu, M. (2018). TAM: Test analysis modules. R package version 2.13-15. Available at: https://cran.r-project.org/package=TAM (Accessed May 18, 2022).

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educ. Meas. Issues Pract.* 24, 3–13. doi: 10.1111/j.1745-3992.2005.00006.x

Rudner, L., and Schafer, W. (2002). *What Teachers Need to Know about Assessment*. Washington, DC: National Education Association.

Sarac, M., and Feinberg, R. A. (2022). "Exploring the Utility of Nonfunctional Distractors" in *The Annual Meeting of the Psychometric Society*, vol. *83–93* (New York: Springer, Cham)

Shinn, M. R., and Shinn, M. M. (2002). *AIMSweb Training Workbook: Administration and Scoring of Reading Maze for Use in General Outcome Measurement*. Eden Prairie, MN: Edformation, Inc.

Simms, L. J., Zelazny, K., Williams, T. F., and Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychol. Assess.* 31, 557–566. doi: 10.1037/pas0000648

Snow, C. (2002). *Reading for Understanding, Toward an R&D Program in Reading Comprehension*. Santa Monica, CA: RAND.

Stecker, P. M., Fuchs, L. S., and Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychol. Schools* 42, 795–819. doi: 10.1002/pits.20113

Stecker, P. M., Lembke, E. S., and Foegen, A. (2008). Using progress-monitoring data to improve instructional decision making. *Prevent. School Failure* 52, 48–58. doi: 10.3200/PSFL.52.2.48-58

Stevens, E. A., Park, S., and Vaughn, S. (2019). A review of summarizing and main idea interventions for struggling readers in grades 3 through 12: 1978–2016. *Remed. Spec. Educ.* 40, 131–149. doi: 10.1177/0741932517749940

Støle, H., Mangen, A., and Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Comput. Educ.* 151:103861. doi: 10.1016/j.compedu.2020.103861

Stuart, M., Stainthorp, R., and Snowling, M. (2008). Literacy as a complex activity: Deconstructing the simple view of reading. *Literacy* 42, 59–66. doi: 10.1111/j.1741-4369.2008.00490.x

Svensson, I., Fälth, L., Tjus, T., Heimann, M., and Gustafson, S. (2019). Two-step tier three interventions for children in grade three with low reading fluency. *J. Res. Spec. Educ. Needs* 19, 3–14. doi: 10.1111/1471-3802.12419

Tolar, T. D., Barth, A. E., Fletcher, J. M., Francis, D. J., and Vaughn, S. (2014). Predicting reading outcomes with progress monitoring slopes among middle grade students. *Learn. Individ. Differ.* 30, 46–57. doi: 10.1016/j.lindif.2013.11.001

Torppa, M., Vasalampi, K., Eklund, K., Sulkunen, S., and Niemi, P. (2020). Reading comprehension difficulty is often distinct from difficulty in reading fluency and accompanied with problems in motivation and school well-being. *Educ. Psychol.* 40, 62–81. doi: 10.1080/01443410.2019.1670334

van Dijk, T. A., and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Cambridge, MA: Academic Press.

Varma, S. (2006). Preliminary item statistics using point biserial correlation and p value. Educational Data System. Available at: https://eddata.com/wp-content/uploads/2015/11/EDS_Point_Biserial.pdf (Accessed May 18, 2022).

Vaughn, S., and Fletcher, J. M. (2012). Response to intervention with secondary school students with reading difficulties. *J. Learn. Disabil.* 45, 244–256. doi: 10.1177/0022219412442157

Weigle, S. C., Yang, W., and Montee, M. (2013). Exploring reading processes in an academic reading test using short-answer questions. *Lang. Assess. Q.* 10, 28–48. doi: 10.1080/15434303.2012.750660

Wijekumar, K., Meyer, B. J., Lei, P., Beerwinkle, A. L., and Joshi, M. (2019). Supplementing teacher knowledge using web-based Intelligent Tutoring System for the Text Structure Strategy to improve content area reading comprehension with fourth- and fifth-grade struggling readers. *Dyslexia* 1–17. doi: 10.1002/dys.1634

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