frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 21 August 2024
DOI 10.3389/fcomm.2024.1384395

OPEN ACCESS

EDITED BY
Andrea Listanti,
University of Cologne, Germany

REVIEWED BY
Alexis Lopez,
Educational Testing Service, United States
Andrea Scibetta,
Foreigners University of Siena, Italy

*CORRESPONDENCE
Fauve De Backer
✉ fauve.debacker@ugent.be

RECEIVED 09 February 2024
ACCEPTED 15 July 2024
PUBLISHED 21 August 2024

CITATION
De Backer F, Vantieghem W, Slembrouck S and Van Avermaet P (2024) The dynamics of multilingual assessment: exploring the impact of linguistic accommodations on science achievement. *Front. Commun.* 9:1384395. doi: 10.3389/fcomm.2024.1384395

The dynamics of multilingual assessment: exploring the impact of linguistic accommodations on science achievement

Fauve De Backer[1]*, Wendelien Vantieghem[1], Stef Slembrouck[2] and Piet Van Avermaet[1]

[1]Centre for Diversity and Learning, Linguistics Department, Ghent University, Ghent, Belgium, [2]Linguistics Department, Ghent University, Ghent, Belgium

This study examines the impact of linguistic accommodations on the science performance of multilingual pupils. In a randomized controlled trial conducted in Flanders (Belgium), pupils aged 9–12 were assigned to one of three conditions: a control group taking a science test in the language of schooling (n = 64), a group receiving a written bilingual test in both the pupils' L1 and language of schooling (n = 64), and a group with a written bilingual test accompanied by read-aloud accommodations in both languages (n = 69). The hypothesis posited that pupils in accommodated conditions would outperform those in non-accommodated conditions. However, univariate analysis of variance did not reveal significant differences between conditions, suggesting that accommodations did not lead to higher test scores. Subsequent multiple linear regression within the condition involving the bilingual test with read-alouds examined how within-group variance impacted accommodation effectiveness, considering both main effects and interaction effects. Results indicate that proficiency in the L1 and frequency of read-aloud use in the L1 significantly predict science performance. Notably, for pupils who frequently used read-alouds, the significant interaction effect with L1 proficiency suggests an amplified beneficial effect on the test scores when pupils are more proficient in their L1.

KEYWORDS

education, multilingualism, multilingual assessment, multilingual testing, accommodations, effectiveness, validity, science achievement

Introduction

Assessing multilingual pupils' competences in content-related areas can be quite a challenge for educators, since the pupils' proficiency in the language of schooling impacts the test results (Menken, 2010; Abedi, 2017). Traditionally, language proficiency and content knowledge are treated as separated constructs while in fact they are much more related than often assumed. For example, pupils' reading scores are a good predictor of their science achievement (O'Reilly and McNamara, 2007). From the perspective of content assessment, language is seen as a source of construct-irrelevant variance – '*variance in scores that is not related to the construct being assessed*' (p. 4) – and vice versa (Abedi, 2004; Llosa, 2017). Hence, pupils that are more proficient in the language of schooling, score higher on tests, even when language proficiency is not the construct intended to be measured. When language proficiency unnecessarily interferes with pupils' ability to illustrate their content-knowledge, this poses a

serious validity concern (Wolf et al., 2012), with validity referring to the quality of the decisions and inferences made based on the scores (Chan, 2014).

Scholars have actively sought solutions to address these challenges, recognizing that diverging from standard procedures may be essential to ensure the validity of test score interpretations (Reynolds et al., 2021). One approach involves adopting assessment accommodations, as outlined by Butler and Stevens (1997, p. 5), who define accommodations as *"support provided to students for a given testing event, either through modification of the test itself or through modifications of the testing procedure, to help students access the content in English and better demonstrate what they know."* According to the Standards (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), the goal of accommodations is to provide the most valid and accurate measurement of the construct for each pupil that is being tested.

There is not yet a consensus in research on which accommodations are valid and effective under which circumstances (Abedi, 2017). If the accommodation succeeds in removing the disadvantage for the multilingual pupil, it is considered effective (Li and Suen, 2012). The goal of the assessment accommodations is reducing the language barrier for multilingual learners (Abedi et al., 2003; Rivera et al., 2006; Kopriva et al., 2007; Pennock-Roman and Rivera, 2011). The effective removal of barriers could possibly improve the ability to measure the construct as intended, hence heightening the validity of the assessment (Kopriva et al., 2007). Examples of accommodations are giving pupils extra time, allowing them to be tested individually or in small groups, providing pupils with translations, glossaries, simplified text (Pitoniak et al., 2009), bilingual tests, a (mono-or bilingual) dictionary, and read-aloud test administration (Abedi et al., 2004).

In the present study we will focus on the effectiveness of bilingual tests and read-aloud accommodations.

Literature review

Translation as an assessment accommodation

When pupils *solely* receive a translation of the assessment, this accommodation may not provide valid results when the languages of instruction and assessment are not aligned (Abedi et al., 2004). For example, when a pupil learns content-area concepts in English but is tested in Spanish, the pupil will lack the content terminology in his first language (L1)[1] to be able to perform adequately on the assessment (Abedi, 2017).

When multilingual pupils (MP) receive the test both in the language of schooling and a translation, this is called a bilingual test or a dual-language test. A systematic understanding of the benefits and limitations of bilingual tests is still lacking (Kieffer et al., 2009; Pennock-Roman and Rivera, 2011; Robinson, 2011; Kieffer et al., 2012; Ong, 2013), but there have been some studies on the topic. In the study by Shohamy (2011), the former USSR pupils who received a bilingual Hebrew-Russian mathematics tests, significantly outperformed pupils who got a monolingual Hebrew version of the test. Duncan et al. (2005) showed that eighth-graders had positive perceptions of the usefulness of bilingual mathematics tests. These results were confirmed by research of De Backer et al. (2019) on the perceptions of fifth-graders on the helpfulness of a science bilingual test. Furthermore, both of these studies (Duncan et al., 2005; De Backer et al., 2019) found that the bilingual test increased pupils' understanding of the test items and/or improved comprehension of specific words. However, pupils themselves indicated that the effectiveness of this accommodation depended on the level of proficiency in their L1. If language skills in the L1 are not well developed or if the content-specific language is missing, a translation might indeed not help (Butler and Stevens, 1997; Kieffer et al., 2009; Robinson, 2011).

Read-aloud accommodations

Although still inconclusive, there have been some studies on the effectiveness of read-aloud accommodations in the language of schooling. While some research has reported on the possible effectiveness of these accommodations (Rivera et al., 2006), other studies found no effects of this kind of accommodations on pupils' test scores (Kopriva et al., 2007; Reed et al., 2014; Abedi, 2017). There is even less experimental research available on the effectiveness of read-aloud accommodations in the pupils' L1. According to Francis et al. (2006), the read-aloud of *test directions* in the pupils' L1s appears to be responsive to their needs. Many multilingual pupils are illiterate in their L1 or they are second or third generation and have not received any formal instruction in the country of origin. Hence, in these cases, a written translation would provide little support to these pupils (Stansfield, 2011). Therefore, read-aloud accommodations could be of more advantage than written support.

Main effects of background characteristics

Research on accommodations for multilingual pupils has only started to emerge within the last decades (see for example Sireci et al., 2003; Kieffer et al., 2009; Wolf et al., 2012; Abedi, 2017; Koran and Kopriva, 2017). Consequently, there is a limited number of studies per accommodation type that has taken into account the impact of pupils' background characteristics/controlled for pupils' background variables (Acosta et al., 2008). Even though such studies are limited, other studies have unveiled the effect of background characteristics on achievement, showing for instance how gender, grade retention and SES are related to science achievement.

The impact of gender on science achievement is somewhat unclear, with results from the Trends in International Mathematics and Science Study (TIMSS) indicating that in Flanders, boys in fourth grade of primary education tend to score higher on science achievement than girls (Gielen et al., 2012), while the Programme for International Student Assessment (PISA) 2015 results show similar levels of science performance for boys and girls (OECD, 2016c).

1 Throughout this text we will use the term L1, referring to the first language pupils learned. We do however acknowledge this term has disadvantages as well, as it does not always reflects current language use.

While a lot of studies (see Jimerson, 2001; Abedi et al., 2003; Kieffer et al., 2009; Goos et al., 2021 for a systematic review) discuss the possible short-term and long-term effects of grade retention, a study suggested that for multilingual pupils specifically grade retention was initially negatively related to science achievement but this association became non-significant when proficiency in the language of schooling and reading performance were taken into account (Van Laere et al., 2014).

Socio-economic status is typically related to academic achievement in general and to science performance in particular. Belgium is among the few high-performing countries in which the relationship between SES and student performance is stronger than average (OECD, 2016c).

Concerning migration status, results from PISA indicate that there is an average difference in science performance between immigrant and non-immigrant students, even after taking their socio-economic status into account (OECD, 2016b). Furthermore, the older children are on arrival, the less well they perform on the reading assessment of PISA at age 15 (OECD, 2012). Arriving at a later age and being unable to speak and read the language of schooling makes pupils more vulnerable than arriving at a younger age.

Differential impact of assessment accommodations by pupils' background characteristics

Acosta et al. (2008) identify the need for a theoretical framework to analyze the linguistic challenges pupils face when they are not yet proficient in the language of schooling. Second language acquisition research can contribute to such a framework. For example, research has already indicated that at the early stages of language acquisition, pupils need more time to encode and decode text in the language of schooling than their native speaking peers (Acosta et al., 2008).

Pupils also differ in the strategies they use for assessment accommodations. Some pupils who receive bilingual tests start reading in the language of schooling and switch to their L1 when they encounter difficult words or sentences, while others start in their L1 right away. There is a huge variety in strategies and use of the assessment accommodations depending on several factors, such as level of language proficiency, personal preferences, and difficulty of the test item (De Backer et al., 2019). In their meta-analysis, Pennock-Roman and Rivera (2011) included 14 US studies with ELLs (English Language Learners) on the effectiveness of accommodations. Their findings showed that for pupils with low proficiency in the language of schooling, translated assessments were found to be effective measures to improve pupils' performance. This study confirms that the effectiveness of bilingual tests as an accommodation depends on both language proficiency in the language of schooling, and on literacy skills as well as content knowledge in the L1 (Francis et al., 2006).

Few accommodations are likely to be effective for all pupils who are not yet proficient in the language of schooling. There is no one-size-fits all, but rather a need to differentiate according to pupils' characteristics. Pupils at the lowest levels of language proficiency in the language of schooling benefit more from oral rather than written accommodations in the language of schooling (Acosta et al., 2008). For pupils who have received instruction in their native language and are literate in their L1, accommodations in their L1 tend to be especially useful. In other words, for literate pupils who just arrived, written translations and bilingual tests are most beneficial. For illiterate newly arrived pupils, read-aloud accommodations are recommended (Acosta et al., 2008).

Since there is so much within-group variance, Elliott et al. (2009) suggest to take the research on assessment accommodations a step further by providing students with several accommodations at once, as providing pupils with a sole accommodation might not be beneficial enough. In the present study, we therefore not only explore the effectiveness of read-aloud and bilingual tests, but also the combination of both in a set.

In sum, there is a need to expand the research on assessment accommodations for multilingual pupils. We know little about the effectiveness of multilingual assessment measures, such as a bilingual test, read-aloud accommodations in pupils' L1s and the combination of both. Moreover, to our knowledge the relation between a set of direct linguistic accommodations and science achievement has not been explored empirically with regard to different pupils' background characteristics.

The present study

To this day, the effectiveness of several possible accommodations for multilingual pupils remains underexplored. What is more, most research on this topic has been done in the United States, where a majority of multilingual pupils tends to be Spanish-speaking. However, in order to ascertain whether accommodations are effective regardless of educational context and language, it is important that this research is supplemented with studies in different contexts. Consequently, this study will assess the effectiveness of bilingual tests and the possible added value that read-aloud accommodations might have in Flanders, the northern part of Belgium. These will be assessed in a sample of multilingual speakers in Flanders (Belgium) by means of a science test. Due to the migration history of Belgium, children of non-western European descent tend to have predominantly Turkish, Moroccan or Eastern-European roots. What is more, pupils with Turkish or Moroccan roots are often second or third generation, whereas students with eastern-European background tend to be more often first generation. Hence, the Flemish context provides us with a research setting in which we can not only observe a diversity of languages present in the classroom, but also one in which the mastery of both the language of schooling (in Flanders, this is Dutch) and the L1 varies depending on the migration history of the family. This gives a unique opportunity to assess the effectiveness of accommodations in a diverse sample.

The Common European Framework of References for Languages (CEFR) distinguishes between multilingualism and plurilingualism and defines the latter as *'the dynamic and developing linguistic repertoire of an individual user/learner'* (Council of Europe, 2001, p. 4). At the same time, The Council of Europe (2020) defines multilingualism as the co-existence of different languages in a given society. Franceschini (2016) defines multilingualism as: "The capacity of societies, institutions, groups, and individuals to engage on a regular basis in space and time with more than one language in everyday life" (p. 33). In Flanders, multilinguals are most often referred to as 'linguistically different' (*anderstaligen*) (Agirdag et al., 2014), a container term which negatively defines speakers in relation to the language which they do not have. Since the research

context of this study is highly diverse in terms of the language use and proficiency of multilingual pupils, we use the term 'multilingual pupils' to refer to a group that spans the entire continuum of language proficiency. This continuum ranges from beginners, who have minimal proficiency in the instructional language, to advanced speakers, who are highly proficient and near-native in their language skills.

Flemish education is guided by the principle of 'freedom of education' (Eurydice, 2020), which results in a wide variety of forms and types of assessments. There is a shared curriculum, but the assessments are mainly developed by the teachers themselves, with Belgium having no tradition yet of nationwide standardized testing (Ysenbaert et al., 2017). There are no central guidelines for the use of accommodations in testing. Hence, individual teachers decide on the accommodations that are allowed. Research in Flanders has shown that the most commonly used accommodations as decided by individual teachers are linguistic modification of the test questions (modifying the language of a text while keeping the content intact, for example by shortening sentences or using familiar or frequently used words), reading questions out loud to pupils, and being more tolerant toward grammatical and spelling mistakes (De Backer et al., 2017).

As described, previous research has indicated important ways in which the effectiveness of accommodations may vary depending on the background characteristics of pupils. Therefore, this study will explicitly explore within-group variance among multilingual students. Consequently, the effectiveness of these accommodations will be assessed in a sample consisting solely of multilingual speakers, where we will explore to what extent the use and effectiveness of the accommodations is influenced by pupils' background characteristics. We will ascertain these effects by considering pupils' science scores. By doing so, we follow previous studies (e.g., Abedi et al., 2003; Kieffer et al., 2009), who have used test scores to establish the effectiveness of accommodations.

In short, this study assesses the following research questions (RQ):

- RQ1: To what extent do bilingual tests impact multilingual learners' achievement on a science assessment?
- RQ2: To what extent does a set of accommodations (consisting of both a bilingual test and read-alouds in the language of schooling and L1) impact multilingual learners' achievement on a science assessment?
- RQ3: How does the impact of the use of accommodations change based on pupils' background characteristics (such as literacy and proficiency in both the language of schooling and L1)?

Analyses on the impact of background characteristics on the use and functioning of accommodations control for possible confounders which previous research has identified as impacting science achievement. These controls include gender (Gielen et al., 2012; OECD, 2016a), socio-economic background (OECD, 2016a), grade retention (Jimerson and Ferguson, 2007; Van Laere et al., 2014), and schooling in the native country (OECD, 2012).

Methods

Participants

Schools were visited throughout 2016–2017. In order to have enough power for the analyses, measures were taken to maximize the number of students with a multilingual background. Consequently, only schools in urban regions in Flanders (Belgium) were selected. Schools were selected to ensure that those with a high proportion of pupils speaking a language at home different from the language of instruction were more frequently included in the sample. Therefore, all primary schools in Flanders were contacted whose population consisted of at least 60% pupils whose L1 is not Dutch. Of the 103 contacted schools, 35 agreed to participate, translating to a response rate of 34%. School principals indicated that their refusal to participate in the study was mostly fueled by either their involvement in other research projects, or the existent heavy workload of the staff. Note that no school refused participation because of the research topic.

This study is part of a larger research project. In the research project, all pupils ($n = 1{,}022$) in the 35 participating classrooms took part in the online survey and science test. All parents of pupils in the fifth grade of the participating schools were asked for their consent to let their child participate in the study. All pupils with parental consent for participation took part in the research project. For this study, we selected all participating Polish and Turkish pupils. We selected these pupils to take into account Belgium's migration history: children of non-Western European descent primarily trace their roots to Turkish, Moroccan, or Eastern European origins. Furthermore, those with Turkish or Moroccan ancestry often belong to second or third generations, while individuals with Eastern European backgrounds are predominantly first-generation immigrants. Hence the current study involved 197 multilingual Turkish and Polish children who attended fifth grade of primary education.

Participants were aged between 9 and 12 years ($M = 10.84$ years, $SD = 0.67$) and the sample consisted approximately equally of boys and girls (49,2% girls; 47,2% boys), illustrating the representativeness of the sample. The majority of pupils (88%) had a migration background: 23% of the pupils was first generation, 39.3% second generation, 30.6% has one parent that is foreign-born and is part of the so-called 2.5 generation, and 4.9% was third generation. Amongst the pupils, 65.6% had no experience with schooling in another country than Belgium, 7.2% had received instruction in another country for less than a year, 16.4% had experience with schooling in another country between 1 and 5 years, and almost 11% took classes for more than 5 years in another country, illustrating the variety in prior schooling experience in the L1.

Descriptive data for the participants are shown in Table 1, including the N, means and standard errors.

Procedure

The experimental research design consists of three research conditions: (1) a control condition consisting of a non-accommodated science test or 'Dutch-only test (DU/A-); (2) intervention arm A consisting of a bilingual science test, providing students with either a Polish or Turkish translation depending on their migration background (BIL/A−); and (3) intervention arm B consisting of a bilingual science test with additional read-aloud options in both the language of schooling and L1 (respectively Turkish or Polish) (BIL/A+). In both intervention arms, the Dutch and translated version of the question were simultaneously displayed on the screen. In the conditions with additional audio-support (i.e., intervention arm B), the pupils could not only see the side-by-side translations, but were

TABLE 1 Descriptive statistics of the participants in the testing conditions.

Variables	Total sample				Non-accommodated				Bil/A–				BIL/A+			
	%	M	SD	N	%	M	SD	N	%	M	SD	N	%	M	SD	N
Science achievement		18.38	5.35	197		18.64	0.65	64		18.42	0.72	64		18.10	0.62	69
Proficiency language of schooling		4.07	0.56	184		4.12	0.52			4.01	0.59			4.06	0.57	
Literacy language of schooling		4.07	0.63	184		4.12	0.63			4.02	0.62			4.06	0.65	
Literacy L1		3.78	1.05	191		3.85	1.13			3.66	1.07			3.82	0.94	
Proficiency L1		4.32	0.77	190		4.47	0.73			4.25	0.78			4.24	0.79	
Education country of origin (Never)	7.1%			197	9.4%			64	7.8%			64	4.3%			69
SES						2.30	1.6			1.89	1.43			1.85	1.51	
Gender (Female)	49.2%				61%				45%				42%			
Grade Retention (Yes)	38.1%				45%				39%				30%			
Use of audio language of schooling														0.05	0.09	
Use of audio L1														0.09	0.18	

also offered the possibility to listen to a read-aloud version of every question and the multiple-choice answers in both L1 and the language of schooling.

Prior to visiting the schools, teachers were asked to fill in a list in order to gather background information of the pupils, including the L1 of the pupils. These alphabetical lists were used by the researcher to randomly assign participants to different research conditions, of course with the pre-condition that only pupils who spoke either Turkish or Polish could be assigned to the bilingual condition.

Randomization was at the level of the individual in 1:1:1 ratio. For example, Turkish pupil 1 = control condition, Turkish pupil 2 = intervention arm A, Turkish pupil 3 = intervention arm B. When the list in class 1 ended for example with intervention arm A, the next class started with intervention arm B and so on.

Since research has shown that unscripted accommodations in the category of oral clarification could lead to variations in test administration, which can create construct-irrelevant variance (Acosta et al., 2008), we opted for pre-recorded read-aloud accommodations by native speakers with standard pronunciation and intonation patterns in all three languages (Dutch, Polish, and Turkish). All pupils were tested during regular class hours. They were not given any time constraints, since previous research suggests that dual language accommodations require generous time limits in order to be effective (Pennock-Roman and Rivera, 2011). Pupils were not familiar with any assessment accommodations nor with taking a test digitally. Video instructions for each research condition on how to take the test were provided to pupils, modeling the accommodations and encouraging pupils to make use of them. In all testing conditions, pupils filled in a background questionnaire and took a science test on the computer through LimeSurvey.

Measures

Science achievement was measured by means of a test that consists of the 43 multiple-choice science items from the released science items for fourth grade of the Trends in International Mathematics and Science Study (TIMSS), which can be considered a standardized, curriculum-based achievement test. These items were chosen to maximize equivalence in content, reliability, difficulty level and validity between the original and the translated versions, since international large-scale assessments such as TIMSS have strict procedures. The TIMSS 2011 items were prepared in English and translated into many other languages for use in participating countries around the world. In translating the tests, every effort was made to ensure that the meaning and difficulty of the items did not change. Translated versions of the science test in Polish and Turkish were made available from the National Project Centers in Poland and Turkey. Scoring guides are provided for constructed response items. The answers on the science test (43 items) were binary coded (1 = correct answer; 0 = incorrect answer). The average test score for the 197 pupils was 18.38 ($SD = 5.35$).

Proficiency in the language of schooling (L2) and in the L1 was measured through a scale in which pupils had to self-assess their proficiency in speaking and listening in both L2 and L1. The *L2 proficiency scale* consists of 5 items, including items such as "When I watch Dutch Television, I understand everything." These items are rated on a 5-point Likert scale (1 = completely disagree; 5 = completely agree). The scale also includes items such as "How well do you speak Dutch," rated on a five-point Likert scale (1 = very poor; 5 = very well). *L1 Proficiency* consists of 2 items: "How well do you speak your first language" and "How well do you understand your first language." These items are rated on a five-point Likert scale (1 = very poor; 5 = very well). The mean sum of scores was used to construct both scales, which displayed good internal reliabilities: $\alpha_{L1proficiency} = 0.83$; $\alpha_{L2proficiency} = 0.74$.

Literacy in the language of schooling and in the L1 was measured in a similar way. Pupils made a self-assessment of their proficiency for reading and writing skills in both language of schooling and L1 on, respectively, 4 and 2 items with a five-point Likert scale. The scale includes items such as "How well do you read in Dutch?," rated on a five-point Likert scale (1 = very poor; 5 = very well) and "Reading a

book in Dutch is easy to me" (1 = completely disagree; 5 = completely agree). Scales were constructed by using mean sum of scores with good internal reliabilities $\alpha_{L1literacy} = 0.89$; $\alpha_{L2literacy} = 0.80$.

Pupils in the testing condition with audio support had the opportunity to listen to both test questions and the answering options, as often as they wanted to. When pupils clicked the audio button, a registration of their click was made. The proportion of clicks (items that were listened to/total available audio support options) was calculated and resulted in two variables to measure the *frequency of use of the read-aloud accommodation* in both the language of schooling ($M = 5\%$; $SD = 0.09$; min. = 0.00 and max. = 0.50); and L1 ($M = 9\%$; $SD = 0.18$; min. = 0.00 and max. = 0.81).

Demographics

There were items on pupils' *gender* (boy = 2, girl = 1) and *grade retention* [no grade retention = 0, repeated one or more year(s) = 1]. The socio-economic position of the pupils was derived from the work situation of the parents. We asked students to fill out the last or current employment of both their father and mother. Employment was attributed a score based on the EGP-classification (Erikson and Goldthorpe, 2002) ranging from 1 (unskilled manual labor) to 8 (high-grade professionals and managers). The child received the highest SES-score of both parents, reflecting the dominant socioeconomic position of the family as a whole (Forehand et al., 1987; Erikson and Goldthorpe, 1992). In this analysis, this measure is used as a continuous variable with a mean score of 1.85 ($SD = 1.51$). Pupils were asked whether they ever had schooling in another country than Belgium and for how long, resulting in a categorical variable: 'Education in the country of origin' (0 = never, 1 = less than 1 year, 2 = 1–5 years, 3 = more than 5 years).

Analysis

Analyses were conducted in IBM SPSS Statistics for Windows, version 24 (IBMCorp, 2016). For RQ1 and 2, which both assess to what extent the (set of) accommodations impact the multilingual learners' science achievement, a univariate analysis of variance (ANOVA) was conducted. This test ascertains whether statistically significant differences exist between the test scores for pupils in the three experimental conditions. The hypothesis was that pupils in the group that received the bilingual test will score higher on the science test than pupils taking a non-accommodated test, but lower than the group who receive a set of accommodations (BIL/A+ > BIL/A− > DU/A−).

For RQ3, which explores within-group variance, multiple regression analyses was used. These analyses consider both main effects and possible interaction-effects, to fully explore how the pupils' linguistic background and their use of accommodations might differentially impact their science achievement, adding in a first and second model the variables related to the, respectively, the language of schooling (i.e., proficiency, literacy) and L1. Then we included the control variables to check whether the associations remained, diminished or disappeared after controlling for first country of schooling and secondly gender, grade retention, etc. Note that we focus on the pupils in the third condition for these analyses, as they are the only ones who had access to the full set of accommodations. However, because of the limited sample size in this group, statistical power must be taken into account. Because of this, marginally significant effects (i.e., $p < 0.10$) are discussed as well, as these might unveil interesting patterns in the data. Secondly, we worked in a step-by-step fashion: first assessing the impact of the main variables, then adding the control variables, and lastly exploring possible interaction-effects. To achieve parsimonious models and preserve statistical power, only significant variables were retained in subsequent analyses. Similarly, when assessing which main effects are significant, continuous variables were used where possible. However, to better interpret interaction effects in the final model, the continuous variable of "frequency of use of the read-aloud accommodation" was categorized into three levels: "no use" (no use at all), "low use" (cut-off at 42%) and "medium to high use" (cut-off at 81%). We dummy-coded these levels, with "no use" as the reference level.

Results

Univariate analysis of variance: research questions 1 and 2

The initial ANOVA across groups showed no significant differences in scores, $F\ (2,194) = 0.17$, $p = 0.84$, $\eta^2 = 0.006$. This preliminary analysis indicates that none of the intervention arms significantly improved the test score of multilingual pupils.

Multiple linear regression: research question 3

Subsequently, a multiple linear regression within the group of pupils who received the set of accommodations (i.e., intervention arm B) was performed to determine how the background characteristics of pupils contributed significantly to their scores. The results of the analyses for the pupils in the BIL/A+ condition are presented in Table 2.

Model 1: language proficiency and literacy in the language of schooling

In the first model, language proficiency and literacy in the language of schooling were added as explanatory variables. However, both language proficiency and literacy in the language of schooling were not significant predictors of science achievement [$F(2,56) = 0.445$, $p = 0.643$, $f^2 = 0.016$]. Consequently, both factors were omitted for the subsequent analyses (Table 3).

Model 2: language proficiency and literacy in the L1

In the second model, language proficiency and literacy in the L1 were added as explanatory variables. Proficiency in the L1 is a significant predictor of science achievement [$F(2,62) = 4.538$, $p < 0.05$, $f^2 = 0.015$] with an R^2 of 0.128. Pupils' science achievement increases with almost 2.5 points for each unit increase of Proficiency in the L1. Literacy in L1 was not a significant predictor of science achievement ($t = X$, $p > 0.010$). Consequently, Literacy L1 was omitted for the subsequent analyses.

TABLE 2 Summary of hierarchical regression analysis for science achievement within intervention arm B ($N = 53$).

	Model 1			Model 2			Model 3			Model 4			Model 5			Model 6		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE(B)*	*β*	*B*	*SE(B)*	*β*	*B*	*SE(B)*	*β*
Constant	18.156	4.632		9.264	3.64		6.144	3.153		2.854	3.956		5.099	3.6		5.960	3.211	
Proficiency language of schooling	−1.384	1.616	−0.166															
Literacy language of schooling	1.273	1.423	0.174															
Literacy L1				−0.467	0.702	−0.086												
Proficiency L1				2.496	0.843	0.383**	3.131	0.750	0.492***	2.405	0.764	0.428**	2.333	0.711	0.414**	2.640	0.731	0.405**
Education country of origin							0.842	0.490	0.203°	0.871	0.504	0.240°	0.457	0.466	0.125			
SES										0.480	0.390	0.164						
Gender										2.113	1.138	0.245°	0.803	1.024	0.093			
Grade Retention										−1.125	1.268	−0.122						
Use of audio language of schooling													4.826	5.929	0.102			
Use of audio L1													9.799	3.581	0.336**	9.718	3.119	0.349**
R^2	0.016			0.13			0.225			0.235			0.339			0.24		

°$p < 0.1$, **$p < 0.01$, ***$p < 0.001$.

10.3389/fcomm.2024.1384395

TABLE 3 Summary (2) of hierarchical regression analysis with interaction effects for science achievement within intervention arm B ($N = 62$).

Variable	Model 1			Model 2		
	B	*SE B*	*β*	*B*	*SE B*	*β*
Constant	7.078	3.210		15.553	4.869	
Proficiency L1	2.276	0.715	0.373**	0.332	1.104	0.054
Use of audio L1						
Low use	0.197	1.398	0.017	−12.524	6.261	−1.091*
Medium to high use	4.193	1.355	0.382**	−8.006	6.670	−0.730
Proficiency L1 × Low use				2.875	1.564	1.088°
Proficiency L1 × Med. to high use				0.726	0.352	1.115*
R^2	0.193			0.301		

°$p < 0.1$, *$p < 0.05$, **$p < 0.01$.

Model 3: education in the country of origin

In the third model, years of education in the country of origin was added to the model which was significant in predicting science achievement achievement [$F(2,61) = 8.853$, $p < 0.001$, $f^2 = 0.29$] with an R^2 of 0.225 and is therefore retained in the subsequent model.

Model 4: SES, gender, and grade retention

In the fourth model, socioeconomic status, gender and grade retention were added to the model. SES and grade retention were omitted for the subsequent analyses since they were non-significant predictors of science achievement [$F(5,50) = 3.075$, $p < 0.05$, $f^2 = 0.31$] with an R^2 of 0.235. Gender was marginally significant and therefore retained in the model. Results indicate that there is a trend for boys to perform significantly better than girls ($p = 0.069$) for the science test. Introducing the control variables did not change the overall significance of the earlier model. Nevertheless, even though proficiency in L1 remained significant, its effect size had diminished.

Model 5: frequency of use of the read-aloud accommodations

In the fifth model, frequency of use of the read-aloud accommodations in both language of schooling and L1 were added to the model. Adding frequency of use of read-aloud accommodations to the model caused Education in the country of origin and Gender to become insignificant in predicting science achievement, while proficiency in L1 remained significant and its effect-size relatively unchanged. Frequency of use of audio in the language of schooling was not a significant predictor, while frequency of audio-use in the L1 was a significant predictor of science achievement [$F(2,62) = 9.811$, $p > 0.001$, $f^2 = 0.31$]. In the final model, we only retained the significant predictors (being Use of audio L1 and L1 Proficiency), and this model has an R^2 of 24%. Pupils' science achievement significantly increases when they are more proficient in their L1 and when they make more frequent use of the audio in their L1.

Interaction effects

In what follows, we test the hypothesis that pupils' use of the accommodation modulates the impact of their linguistic background on science achievement. Since previous models showed that proficiency in L1 and use of the audio in L1 have a significant impact on science achievement, we explore specifically the existence of an interaction effect between these variables. In order to assess this possible interaction effect in a detailed way, we employ the variable on the use of audio in L1 in a categorical fashion. In the first step, we see that the main effect of proficiency of L1 remains positive and significant. For use of the read-aloud in L1, we find a positive significant difference for students who used this accommodation intensively ("Medium to high use") versus those who did not ("No use") ($\gamma = 4.20$; $p < 0.010$), whereas the difference between those who used the accommodation only minimally ("Low use") versus not at all ("No use") is insignificant ($\gamma = 0.20$; $p > 0.100$). This indicates that minimal use of the read-alouds is not enough to obtain higher scores on the science test, while frequent use of the read-aloud does lead to higher science scores.

In the second step, we add the interaction effect. Interestingly, the main effect of language proficiency in L1 disappears ($\gamma = 0.33$; $p > 0.01$). Furthermore, the main effect of frequency of audio-use in L1 changes profoundly: the effect is now significantly negative for students who use the read-aloud minimally when compared to students who do not use the read-aloud ($\gamma = -12.52$; $p < 0.05$). At the same time, there are no significant differences on science achievement between those who use the read-alouds intensively versus those who do not ($\gamma = -8.006$; $p > 0.01$). Factoring in the effect of the interaction-effects, shows us that when students' language proficiency is high, their scores on science increase significantly for students who use the audio-support intensively versus those who do not use the audio-support at all ($\gamma = 0.73$; $p < 0.05$). Hence, the science scores of students who use the audio-support intensively, while initially similar to students who did not use the audio-support at all, are increasingly higher when their proficiency in L1 is better.

The interaction-term between language proficiency in L1 and the minimal use of the read-aloud is positive, but marginally significant ($\gamma = 2.88$; $p < 0.100$). Hence, the science scores of students who use the audio-support minimally is initially far below that of students who did not use the audio-support at all, due to the negative main effect of use of the read-aloud. However, proficiency in L1 does have a positive effect on this association, which in the end manages to compensate for this negative main effect of use of the read-aloud. So, when we compare students with a maximal level of proficiency in L1, students who used the read-aloud minimally do obtain a somewhat

higher score on the science test than those who did not use the accommodation at all.

Discussion and conclusion

In this study, we addressed the following questions: (1) To what extent do bilingual tests impact multilingual learners' achievement on a science assessment? (2) To what extent does a set of accommodations (consisting of both a bilingual test and read-alouds in the language of schooling and L1) impact multilingual learners' achievement on a science assessment? and (3) How does the impact of the use of accommodations change based on pupils' background characteristics (such as literacy and proficiency in both the language of schooling and L1)?

In this randomised control trial, pupils were randomly assigned to different conditions, a control condition without accommodations, an intervention arm (A) in which pupils received a written bilingual science test, and an intervention arm (B) in which they received a set of three accommodations: the bilingual science test in combination with read-alouds in both the L1 and the language of schooling. In this experiment, only multilingual pupils (Turkish and Polish) were considered. Rather than assessing the impact of accommodations by comparing pupils who speak the language of schooling at home to multilingual speakers, as is often done (see for instance Abedi, 2017), this study focused on the within-group differences *among* multilingual speakers. This way, we unveiled the differential way accommodations function for this group and highlight the conditions under which some accommodations may be more beneficial than others for specific students.

RQ 1 and 2: the impact of accommodations on multilingual learners' science achievement

The univariate analysis of variance (ANOVA) showed no significant differences in science achievement between the research conditions. Although this outcome contradicted our initial hypotheses, it is not entirely unexpected, given the cautionary note from Pennock-Roman and Rivera (2012), highlighting the heterogeneous nature of multilingual learners and their varying language skills. This could cause near zero effect sizes because the positive effects for one subgroup could be canceled by the negative effects for pupils who know the language of schooling. Indeed, our detailed analysis of the way in which the set of accommodation impacted science achievement (discussed below) confirms exactly this process. More specifically, our results suggest that depending on language proficiency and frequency of use of the accommodations, some accommodations might become either an asset or a barrier. Another possible explanation for the results is the fact that students' science competences seem to be rather low, with an average test score of 43% ($M = 18.38$; $SD = 5.35$). Linguistic accommodations may function differently with students who are more competent in the content being measured than students with lower skills in the subject. Furthermore, the possible influence of content knowledge also raises questions regarding the way impact of assessment accommodations is usually investigated. That is, assessment accommodations for language learners are often evaluated in terms of effectiveness (the extent to which pupils' test scores improve) and validity [the idea that an accommodation should not affect the test score of pupils who do not need it (Kieffer et al., 2009)]. Most studies on assessment accommodation have not focused on accessibility (Elliott et al., 2009), while the main goal of assessment accommodations is *to help students access the content* (Butler and Stevens, 1997, p. 5). Hence, it is important to keep in mind that the non-significant results from the ANOVA-analyses of this study do not necessarily mean that the accommodations do no succeed in increasing access to certain groups of multilingual pupils. That is, the accommodations might help students access the test questions without this resulting in higher test scores due to for instance low content knowledge of the student. Indeed, from pupils' reports we know that they perceive the bilingual test as helpful even though they do not necessarily believe this would increase their test score (De Backer et al., 2019). As Rios et al. (2020, p. 73) recommend: *the field should shift from asking, "Is a particular accommodation effective?" to "For whom, and under what conditions, is a particular accommodation effective?"*

RQ 3: the impact of pupils' background on the frequency of use and efficacy of accommodations

A multiple linear regression within the group of pupils who received the set of bilingual accommodations (BIL/A+) (intervention arm B) was performed to discover possible differential effects of frequency of use of accommodations based on background characteristics. Not surprisingly, proficiency in the L1 was a significant predictor of science achievement. The more proficient pupils are in their L1, the more beneficial the bilingual accommodations can be. This is in line with pupils' own reports, where they indicated that the effectiveness of the accommodation depends on the level of proficiency in their L1 (De Backer et al., 2019). This aligns with prior research indicating that the effectiveness of translation or read-aloud in the first language (L1) may be compromised if the language skills in L1 are underdeveloped or if the domain-specific language is absent (Butler and Stevens, 1997; Kieffer et al., 2009; Robinson, 2011). In the study of Francis et al. (2006), it was reported that the effectiveness of bilingual tests depends on both language proficiency in the language of schooling and on literacy skills and content knowledge in the native language. Pennock-Roman and Rivera (2012) also report in their meta-analysis that the effectiveness of bilingual tests is sensitive to proficiency in the language of schooling and to literacy skills in the native language. In this study, it was notable that the L1 proficiency rather than literacy skills predicted the science achievement, since one of the accommodations that was provided was read-aloud in the L1.

A second significant predictor of science achievement was the frequency of which pupils use the read-aloud in the L1. The more frequent they use the audio-support, the higher their score on the science test. Quite interestingly, the direct effects of both proficiency in L1 and the use of the L1 audio-support change in important ways when we consider their interaction-effect. That is, our analyses show that the effect of proficiency in L1 is entirely dependent on the extent to which pupils used the read-alouds. That is, for those who did not use the read-alouds at all, the impact of their proficiency in L1 became insignificant. Of course, this is a logical finding considering that a

pupils' level of understanding of a language becomes irrelevant if they do not make use of audio-support in this language.

For pupils who do use the read-aloud accommodation, we note substantial differences between those who use it often and those who used it only in a limited way. For those who used the read-alouds often, we found a significant interaction-effect with L1 proficiency, indicating that the beneficial effect of using the audio-support in L1 on science achievement becomes more pronounced when pupils are more proficient in their L1. Due to this effect, pupils with high L1-proficiency who used the read-alouds frequently tended to outscore students with a similar L1-proficiency who did not use the read-alouds at all. Clearly, these findings highlight the potential beneficial effect of including read-aloud accommodations on tests.

Nevertheless, it is interesting to also take a closer look at students who used the read-alouds only to a limited extent. The interaction effect reveals that students with high proficiency in their first language (L1) only achieved slightly higher scores, and the difference was marginally significant when compared to students with similar L1 proficiency who did not utilize read-alouds at all. This can be attributed to the fact that employing read-alouds in a restricted manner, as opposed to not using them at all, appears to hinder performance on the science test. For students who utilized read-alouds sparingly, this accommodation seemed to act more as a distraction than a support. That is, we uncovered a negative effect on pupils' science achievement of using the read-alouds in a limited fashion versus not at all, that pupils' language proficiency in L1 could only compensate for at high levels. Consequently, students with average or low proficiency in L1 who made little use of the audio-support actually scored lower on the science test than similarly L1-proficient pupils who did not use the audio-support at all. Possible explanations for this negative effect of using accommodations is that in some cases, read-alouds might work as a distraction to the task at hand or might heighten confusion among pupils.

Implications for the research field

The field of multilingual assessment is rather young and many questions remain unexplored. To move forward as a field, we would like to make some suggestions. We believe in the strengths of a mixed-methods approach to gain a comprehensive understanding of the effectiveness of accommodations. Combining quantitative data with qualitative insights from stakeholders, such as educators and pupils, will help ensure the relevance and practicality of different strategies. Specifically, cognitive interviews or think-alouds can provide insight into pupils' cognitive processes when using assessment accommodations and allows researchers to uncover potential barriers. Considering the significant role of familiarity with assessment accommodations, we recommend longitudinal studies to comprehensively assess their effectiveness over time.

Furthermore, a central take-away from the current study is the observation that multilingual pupils are highly diverse in terms of their language use and proficiency, and consequently, we included pupils at all stages of the multilingualism-continuum. This study has drawn attention to the fact that these individual characteristics of multilingual pupils are much more heterogeneous than tends to be commonly assumed, by determining how within-group variance differentially impacts the effectiveness of accommodations. Multilingual learners are often treated as one group with similar characteristics, which our study showed to not to be the case. If we want to develop fair assessments for multilingual learners and thus be responsive to their needs, we need to develop a better and more sophisticated view on what these characteristics are and what works for which pupils.

Additionally, this research has concentrated on the implementation of assessment accommodations for multilingual learners. Despite numerous studies advocating for the advantages of accommodations, there are inherent limitations and criticisms associated with their use. Firstly, the assumption underlying accommodations is that language and content can be separated, and that pupils will be able to show their competences on the content if their limited linguistic abilities do not get in the way (Llosa, 2017). However, the distinction between content and language is, in one sense, artificial, as language and content are inherently interconnected: there is no content without language and there is no language without content. Consequently, the pursuit of a testing environment devoid of construct-irrelevant variance may appear idealistic. Nevertheless, this should not deter efforts to progress toward greater equity in the assessment of multilingual students (Heugh et al., 2017). One step forward in doing so would be to anticipate for linguistic difficulties in the design of tests (Faulkner-Bond and Sireci, 2015), rather than using assessment accommodations as an *ad hoc* solution.

Implications for practice

This study offers valuable recommendations for practice. According to Shafer Willner and Mokhtari (2018), tools and accommodations should be integrated into daily instruction to familiarize students with their usage. Nevertheless, incorporating translations for every test situation in all languages used by students during daily classroom assessments may not be feasible. Achieving construct equivalence is not only technically challenging but also time-consuming and costly, as highlighted by Abedi et al. (2004). Nonetheless, technology is rapidly advancing and machine translations have improved significantly over time. Other possibilities for classroom assessments are the use of glossaries and dictionaries.

In the case of large-scale assessment research and standardized testing, providing translations and/or read-alouds is something that definitely should be considered as our results indicate that these are especially effective for pupils with high proficiency in L1, and hence, might be especially beneficial to a vulnerable group such as newly arrived immigrants.

Besides feasibility, another issue is that few accommodations are likely to be effective for all pupils. This study confirms that there is no one-size-fits-all (Acosta et al., 2008). Instead, there is a necessity to tailor approaches based on individual pupil characteristics. Accommodations in pupils' first language can be especially relevant for pupils' with higher proficiency levels in their L1. At the same time, the results suggest the need for students to be able to familiarize themselves with the available accommodations. With guidance from their teacher, they can figure out which ones are supportive or hindering to them, rather than having to find out at the moment of assessment itself.

Limitations and suggestions for future research

In this study, there was no intervention arm in which pupils received the written test in the language of schooling and the read-aloud in the L1 to explore the unique effect of this accommodation. We only explored the effect of the read-aloud condition in both the language of schooling and the L1 in combination with the bilingual test. It would be worth exploring whether read-alouds in the L1 are best offered in combination with a written translation or as a 'standalone' accommodation. Another limitation is that language proficiency and literacy skills have been measured only via self-report. While it is common practice in educational research to rely on self-reported data, the accuracy of this type of data is open to debate. The meta-study of Kuncel et al. (2005) indicated that self-reported data is generally accurate and can be used as a measure of student achievement. Nevertheless, critical voices must be noted too. For example, Rosen et al. (2017) caution that especially lower-performing pupils tend to overestimate their grades. We tested science performance but relied on self-reported data as measures for pupils' language proficiency and literacy in order to not overload these young pupils with too many tests. Checking the self-reported data with their teachers or standardized language tests could have provided a fuller picture of pupils' language proficiency. Furthermore, we acknowledge that children's science competence was not controlled for. Children were randomly distributed across testing conditions, with one group possibly performing better than another which might have affected the results of this study.

In their book, Melo-Pfeifer and Olivier (2023) address the issue of plurilingual competence in 15 contributions, with one of them focused on translanguaging in assessment (Ascenzi-Moreno et al., 2023) a rather far-reaching approach (Melo-Pfeifer and Olivier, 2023). In this study, we focused on the effectiveness of accommodations for science achievement. Future research could investigate the potential effects of linguistic accommodations on both language proficiency in the language of instruction and pupils' first languages.

Also, looking at test scores is only one aspect of the possible advantages of accommodations. Other aspects are for example increased self-efficacy or well-being, as students feel recognized and accepted as a linguistic minority. This would be interesting to explore in future research, as it is suggested that the acceptance and use of the children's home languages in classroom interactions resulted in an increase in pupils' well-being (Ramaut et al., 2013; Slembrouck et al., 2018). Consequently, it would be worthwhile to explore the effect of exploiting L1s in assessment on the wellbeing and self-efficacy of pupils.

Future research could delve into the intriguing aspect of why the adverse impact is evident among students who utilized read-alouds in a limited manner, while students who employed read-alouds intensively do not experience this effect. When compared to pupils who did not use read-alouds at all or used them extensively, it is possible that these students are more prone to distraction, have difficulty maintaining focus for extended periods, or are more susceptible to feeling overwhelmed by substantial amounts of information. Especially important to note with regards to these findings is the limited use of the read-aloud accommodation in our study for both language of schooling ($M = 5\%$; $SD = 0.09$) and L1 ($M = 9\%$; $SD = 0.18$). Limited utilization not only offers a potential explanation for the absence of a significant difference between groups in the ANOVA analyses but also suggests a lack of familiarity among students with such accommodations. Despite video instructions aiming to encourage accommodation use, this may not have been sufficiently effective. While some accommodations, like dictionaries, require prior experience for optimal utilization, Acosta et al. (2008) argue that this is not necessarily true for others such as oral translations or read-alouds. The findings of the current study, however, cast doubt on this assertion. It appears plausible that students should be acquainted with all accommodations at the time of assessment to maximize their effectiveness. The punitive effect observed for students with limited L1-proficiency underscores the importance of affording every student the opportunity to discern which accommodations might be supportive or detrimental to their unique learning process. Future research could investigate whether multilingual assessment accommodations yield greater efficacy when students are familiar with incorporating their entire linguistic repertoire in daily classroom practices.

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 humans were approved by Ethics committee – faculty of arts and philosophy. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

FB: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. WV: Formal analysis, Supervision, Writing – review & editing. SS: Funding acquisition, Supervision, Writing – review & editing. PA: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The research reported in this article was funded by the Flemish Research Foundation [grant number FWOOPR2015003901]. Uitgegeven met steun van de Universitaire Stichting van België - Published with the support of the University Foundation of Belgium.

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

References

Abedi, J. (2004). The no child left behind act and English language learners: assessment and accountability issues. *Educ. Res.* 33, 4–14. doi: 10.3102/0013189X033001004

Abedi, J. (2017). "Utilizing accommodations in assessment" in Language testing and assessment, encyclopedia of language and education. ed. E. Shohamy (MA Springer: Boston), 2462–2478.

Abedi, J., Courtney, M., and Leon, S. (2003). Effectiveness and validity of accommodation for English language learners in large-scale assessments. Los Angeles: Center for the Study of Evaluation National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C. H., and Lord, C. (2004). Assessment accommodations for English language learners: implications for policy-based empirical research. *Rev. Educ. Res.* 74, 1–28. doi: 10.3102/00346543074001001

Acosta, B. D., Rivera, C., and Shafer Willner, L. (2008). Best practices in state assessment policies for accommodating English language learners: a Delphi study. Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

Agirdag, O., Jordens, K., and Van Houtte, M. (2014). Speaking Turkish in Belgian schools: teacher beliefs versus effective consequences. *Bilig* 70, 07–28. doi: 10.12995/bilig.2014.7001

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (Ed.) (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Ascenzi-Moreno, L., Garcia, O., and Lopez, A. A. (2023). Latinx bilingual students' translanguaging and assessment: a unitary approach. In S. Melo-Pfeifer and C. Ollivier. (Eds.). (2023). Assessment of plurilingual competence and plurilingual learners in educational settings: educative issues and empirical approaches. London: Routledge, 2023, 48–61.

Butler, F. A., and Stevens, R. (1997). Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations (vol. 448). Los Angeles: Center for Research on Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California.

Chan, E. K. H. (2014). "Standards and guidelines for validation practices: development and evaluation of measurement instruments" in Validity and validation in social, behavioral and health sciences. eds. B. D. Zumbo and E. K. H. Chan, vol. *54* (Swizterland: Springer International Publishing), 9–24.

Council of Europe (2001). Common European framework of reference for languages: learning, teaching, assessment. Strasbourg: Council of Europe.

Council of Europe (2020). "Common European framework of reference for languages: learning, teaching, assessment" in Companion volume with new descriptors (strasbourg: Council of Europe).

De Backer, F., Baele, J., Van Avermaet, P., and Slembrouck, S. (2019). Pupils' perceptions on accommodations in multilingual assessment of science. *Lang. Assess. Q.* 16, 426–446. doi: 10.1080/15434303.2019.1666847

De Backer, F., Van Avermaet, P., and Slembrouck, S. (2017). Schools as laboratories for exploring multilingual assessment policies and practices. *Lang. Educ.* 31, 217–230. doi: 10.1080/09500782.2016.1261896

Duncan, T. G., Parent, L. D. R., Chen, W. H., Ferrara, S., Johnson, E., Oppler, S., et al. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Appl. Meas. Educ.* 18, 129–161. doi: 10.1207/s15324818ame1802_1

Elliott, S. N., Kratochwill, T. R., McKevitt, B. C., and Malecki, C. K. (2009). The effects and perceived consequences of testing accommodations on math and science performance assessments. *Sch. Psychol. Q.* 24, 224–239. doi: 10.1037/a0018000

Erikson, R., and Goldthorpe, J. (1992). The constant flux: a study of class mobility in industrial countries. New York: Oxford University Press.

Erikson, R., and Goldthorpe, J. K. (2002). Intergenerational inequality: a sociological perspective. *J. Econ. Perspect.* 16, 31–44.

Eurydice. (2020). Belgium – Flemish community overview. Available at: https://eacea.ec.europa.eu/national-policies/eurydice/content/belgium-flemish-community_en

Faulkner-Bond, M., and Sireci, S. G. (2015). Validity issues in assessing linguistic minorities. *Int. J. Test.* 15, 114–135. doi: 10.1080/15305058.2014.974763

Forehand, R., Middletin, K, and Long, N. (1987). Adolescent functioning as a consequence of recent parental divorce and the parental-adolescent relationship. *J. Appl. Dev. Psychol.* 8, 305–315.

Franceschini, R. (2016). "Multilingualism research" in The Cambridge handbook of linguistic multi-competence. Cambridge handbooks in language and linguistics. eds. V. Cook and L. Wei (Cambridge: Cambridge University Press), 97–124.

Francis, D. J., Rivera, C., Lesaux, N. K., Kieffer, M. J., and Rivera, H. (2006). Practical guidelines for the education of English language learners: research-based recommendation for the use of accommodations in large-scale assessments. Portsmouth, NH: RMC Research Corporation, Center on Instruction. Available at: http://www.centeroninstruction.org/files/ELL3-Assessments.pdf.

Gielen, S., Bellens, K., Belfi, B., and Van Damme, J. (2012). Het vierde leerjaar basisonderwijs in Vlaanderen: Resultaten van TIMSS 2011 in internationaal perspectief en in vergelijking met TIMSS 2003. Leuven: Centrum voor Onderwijseffectiviteit en -evaluatie.

Goos, M., Pipa, J., and Peixoto, F. (2021). Effectiveness of grade retention: a systematic review and meta-analysis. *Educ. Res. Rev.* 34:100401. doi: 10.1016/j.edurev.2021.100401

Heugh, K., Prinsloo, C., Makgamatha, M., Diedericks, G., and Winnaar, L. (2017). Multilingualism(s) and system-wide assessment: a southern perspective. *Lang. Educ.* 31, 197–216. doi: 10.1080/09500782.2016.1261894

IBMCorp (2016). IBM SPSS statistics for windows (version 24.0). Armonk, NY: IBM Corp.

Jimerson, S. R. (2001). Meta-analysis of grade retention research: implications for practice in the 21st century. *Sch. Psychol. Rev.* 30, 420–437. doi: 10.1080/02796015.2001.12086124

Jimerson, S. R., and Ferguson, P. (2007). A longitudinal study of grade retention: academic and behavioral outcomes of retained students through adolescence. *Sch. Psychol. Q.* 22, 314–339. doi: 10.1037/1045-3830.22.3.314

Kieffer, M. J., Lesaux, N. K., Rivera, M., and Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: a meta-analysis on effectiveness and validity. *Rev. Educ. Res.* 79, 1168–1201. doi: 10.3102/0034654309332490

Kieffer, M. J., Rivera, M., and Francis, D. J. (2012). Practical guidelines for the education of English language learners: research-based recommendations for the use of accommodations in large-scale assessments. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C., and Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educ. Meas. Issues Pract.* 26, 11–20. doi: 10.1111/j.1745-3992.2007.00097.x

Koran, J., and Kopriva, R. J. (2017). Framing appropriate accommodations in terms of individual need: examining the fit of four approaches to selecting test accommodations of english language learners. *Appl. Meas. Educ.* 30, 71–81. doi: 10.1080/08957347.2016.1243539

Kuncel, N. R., Credé, M., and Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: a meta-analysis and review of the literature. *Rev. Educ. Res.* 75, 63–82. doi: 10.3102/00346543075001063

Li, H., and Suen, H. K. (2012). Are test accommodations for English language learners fair? *Lang. Assess. Q.* 9, 293–309. doi: 10.1080/15434303.2011.653843

Llosa, L. (2017). Assessing Students' Content Knowledge and Language Proficiency, in *Language Testing and Assessment. Encyclopedia of Language and Education*. Eds. E. Shohamy, I. Or, S. May (Cham: Springer).

Melo-Pfeifer, S., and Ollivier, C. (2023). Assessment of Plurilingual competence and Plurilingual learners in educational settings: educative issues and empirical approaches. *1st* Edn. London: Routledge.

Menken, K. (2010). NCLB and English language learners: challenges and consequences. *Theory Pract.* 49, 121–128. doi: 10.1080/00405841003626619

O'Reilly, T., and McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *Am. Educ. Res. J.* 44, 161–196. doi: 10.3102/0002831206298171

OECD. (2012). Untapped skills. Realising the potential of immigrant students. Ramaut: Linguistics Department, Ghent University.

OECD (2016a). Education at a glance 2016. Paris: OECD Publishing.

OECD. (2016b). Immigrant background, student performance and students' attitudes towards science. Paris: OECD publishing.

OECD (2016c). PISA 2015 results (volume I): excellence and equity in education. Paris: OECD Publishing.

Ong, S. L. (2013). Usefulness of dual-language science test for bilingual learners. *Stud. Educ. Eval.* 39, 82–89. doi: 10.1016/j.stueduc.2012.12.001

Pennock-Roman, M., and Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: a meta-analysis of experimental studies. *Educ. Meas. Issues Pract.* 30, 10–28. doi: 10.1111/j.1745-3992.2011.00207.x

Pennock-Roman, M., and Rivera, C. (2012). Summary on literature on empirical studies of the validity and effectiveness of test accommodations for ELLs: 2005-2012.

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., and Ginsburgh, M. (2009). Guidelines for the assessment of English language learners. Princeton, New Jersey: Educational Testing Service.

Ramaut, G., Sierens, S., Bultynck, K., Van Avermaet, P., Van Gorp, K., Slembrouck, S., et al. (2013). Evaluatieonderzoek van het project 'Thuistaal in onderwijs' (2009-2012): Eindrapport (onuitgegeven onderzoeksrapport). Ghent: Linguistics Department, Ghent University.

Reed, D. K., Swanson, E., Petscher, Y., and Vaughn, S. (2014). The effects of teacher read-Alouds and student silent Reading on predominantly bilingual high school Seniors' learning and retention of social studies content. *Read. Writ.* 27, 1119–1140. doi: 10.1007/s11145-013-9478-8

Reynolds, C. R., Altmann, R. A., and Allen, D. N. (2021). "Assessment Accommodations" in Mastering modern psychological testing (Cham: Springer).

Rios, J. A., Ihlenfeldt, S. D., and Chavez, C. (2020). Are accommodations for English learners on state accountability assessments evidence-based? A multistudy systematic review and Meta-analysis. *Educ. Meas. Issues Pract.* 39, 65–75. doi: 10.1111/emip.12337

Rivera, C., Collum, E., Shafer Willner, L., and Sia, J. K. Jr. (2006). "An analysis or state assessment policies in addressing the accommodation of English language learners" in State assessment policy and practice for English language learners: a national perspective. eds. C. Rivera and E. Collum (Mahwah, NJ: Lawrence Erlbaum), 1–173.

Robinson, J. P. (2011). The effects of test translation on Young English learners' mathematics performance. *Educ. Res.* 39, 582–590. doi: 10.3102/0013189X10389811

Rosen, J. A., Porter, S. R., and Rogers, J. (2017). Understanding student self-reports of academic performance and course-taking behavior. *AERA Open* 3:233285841771142. doi: 10.1177/2332858417711427

Shafer Willner, L., and Mokhtari, K. (2018). Improving meaningful use of accommodations by multilingual learners. *Read. Teach.* 71, 431–439. doi: 10.1002/trtr.1637

Shohamy, E. (2011). Assessing multilingual competencies: adopting construct valid assessment policies. *Mod. Lang. J.* 95, 418–429. doi: 10.1111/j.1540-4781.2011.01210.x

Sireci, S., Li, S., and Scarpati, S. (2003). The Effects of Test Accommodation on Test Performance: A Review of the Literature. Commissioned paper by the National Academy of Sciences/National Research Council's Board on Testing and Assessment.

Slembrouck, S., Van Avermaet, P., and Van Gorp, K. (2018). Strategies of multilingualism in education for minority children. In AvermaetP. Van, S. Slembrouck, GorpK. Van, S. Sierens and K. Marijns (Eds.), The multilingual edge of education. United Kingdom: Palgrave Macmillan.

Stansfield, C. W. (2011). Oral translation as a test accommodation for ELLs. *Lang. Test.* 28, 401–416. doi: 10.1177/0265532211404191

Van Laere, E., Aesaert, K., and van Braak, J. (2014). The role of students' home language in science achievement: a multilevel approach. *Int. J. Sci. Educ.* 36, 2772–2794. doi: 10.1080/09500693.2014.936327

Wolf, M., Kim, J., and Kao, J. (2012). The effects of glossary and read-aloud accommodations on English language learners' performance on a mathematics assessment. *Appl. Meas. Educ.* 25, 347–374. doi: 10.1080/08957347.2012.714693

Ysenbaert, J., Van Avermaet, P., and Van Houtte, M. (2017). Literatuurstudie Evaluatie en diversiteit. Ghent: Steunpunt Onderwijsonderzoek.