The assessment of mathematical literacy of linguistic minority students: results of a multi-method investigation

Wolff-Michael Roth¹, Kadriye Ercikan², Marielle Simon³, Romeo Fola³
¹University of Victoria, ²University of British Columbia, ³University of Ottawa

Abstract Assessing mathematical literacy of students who have limited proficiency in the language of the test is a critical challenge in mathematics education. Previous research indicates that knowledge and competencies of such students are underestimated. This presents a major validity and fairness problem for assessment. Most efforts addressing fairness and validity issues in assessment of linguistic minority students focus on the test language only. To overcome limitations of single approaches, we examine in this study the interaction between the test language and the student language background by means of multiple methods. Thus, we investigate possible linguistic bias of items flagged as functioning differentially (the result of DIF analyses) by means of (a) two levels of expert analyses and (b) student think-aloud protocols to investigate language effects in published mathematics items from the 2000 and 2003 Programme for International Student Assessment (PISA) administration for students attending French schools in Canada and speaking either French or other languages at home. DIF analyses were conducted to identify items on which students from different home language backgrounds attending French schools achieve differently. The expert panels tended to identify surface characteristics of language that may be responsible for group differences but not for the differential effects detected by differential item functioning (DIF). Student think-aloud protocols in part confirm and in part contradict DIF results, providing insights for the source of the differences. Suggestions are provided for further study.

Keywords  Mathematics achievement • linguistic minority students • differential item functioning • expert analyses • mathematical cognition • language differences • think-aloud protocols

1. Introduction

Every three years, the results of the most recent international survey Programme for International Student Assessment (PISA) are both eagerly awaited and dreaded. While this study was conducted, one of the leading national newspapers in Canada calls attention using the title “Canada’s fall in math-education ranking sets off alarm bells” (Alphonso, 2013). The article not only reports that Canada has dropped out of the top 10 "placing 13th overall, down three spots from
2009 and six spots from 2006.” The article further suggests that mathematics scores were dropping not only in Canada but also worldwide and that “the math curriculum, ushered in over the past decade, is to blame for lower scores because it places more emphasis on real-world concepts rather than abstract thinking and practice.” This Canadian reaction is reflected in many other industrialized (Organisation of Economic Co-operation and Development, OECD) countries as well, where mathematics education is either criticized or, if the country has improved in the rankings (e.g., Germany), the distance to the achievement scores of the leading countries is noted as a major concern (e.g., Friedman & Trenkamp, 2013). In France, the most recent results constituted a shock, and “poor showings” have led to reflections concerning the inequities of an educational system where the proportion of elite mathematics students remained the same over the past few years, but where the proportion of students in difficulties (e.g., from disadvantaged families) has risen by a factor of 1.35 (Battaglia & Collas, 2013). Even though there are cautions about the over-interpretation of PISA and other international assessments in education generally (e.g., Ercikan, Roth, & Asil, in press) and in mathematics education specifically (Brown, 2012; Tsatsaroni & Evans, 2013), their influence on national mathematics education policy is considerable (e.g., Boesen et al., 2014) and therefore requires continued scholarly engagement. This is so because the test results “are employed to make significant resource allocations, curriculum planning, and strategic decisions” (Roth, Oliveri, Sandilands, Lyons-Thomas, & Ercikan, 2013, p. 547).

Although the results of the latest PISA assessments tend to make headlines, questions concerning the validity and limitations of the inferences based on it are much more rarely discussed (e.g., Leung, 2014), including, for example, whether multi-lingual tests comparably assess competencies in different language groups—e.g., French and English version of a pan-Canadian test (Ercikan, Roth et al., in press)—or in the same language but for students living in minority situations (Ercikan, Chen et al., in press) or English language learners contexts (e.g., Ercikan et al., 2014). Thus, one recent study has shown that 19, 21, and 25 percent of the reading, mathematics, and science PISA items had different psychometric properties for students from English and non-English language backgrounds attending English-language schools in Australia, Canada, the UK, and the US (Ercikan, Chen, et al., in press). Another study compared PISA mathematics achievement of Quebec students—who speak French at home, live in a French majority setting, and attend schools with French as instructional language—and Ontario students, who attend French-instruction schools in a minority setting and speak (OFF) or do not speak French at home (OFNF) (e.g., Ercikan, Roth, Simon, Sandilands, & Lyons-Thomas, in press). That study showed that of 163 mathematics items across the three assessments conducted between 2000 and 2006, 22 items (13.5%) exhibited bias against one or another group.

---

1 The relevant level differs according to jurisdiction. In Canada and Germany, cultural authority and educational decision-making lies with the provinces (Länder). In countries with central governments, such as France, educational decisions are made nation wide.
The purpose of this research is to investigate possible linguistic bias in PISA items differentially solved by students from different language groups. Most efforts addressing fairness and validity issues in assessments of linguistic minority students focus on assessment language only (Au, 2013; Abedi, 2004; Abedi, Hofstetter, & Lord, 2004; Abedi & Lord, 2001; Butler, Bailey, Stevens, Huang, & Lord, 2004; Nguyen & Cortes, 2013; Vale et al., 2013). In this research we use a mixed-method approach to examine the interaction between the test language and the student language background on large-scale, international examinations with a particular focus on the released PISA 2000 and 2003 mathematics items using four levels of analyses: (a) item-response theory (IRT) based analyses of differential item functioning (DIF) for the selection of items included in analyses (b) through (d), (b) analyses by experts in educational measurement and mathematical cognition, (c) language and curriculum experts, and (d) think-aloud protocols (TAPs) with 33 students from the target populations. DIF analyses help us identify items where students from different language backgrounds have differential response patterns. But this statistical approach does not tell us whether and how language may be at play in the differential response patterns. The expert reviews are the first step to understand if item language may be at play. TAPs help us examine if and how student language may be the source of DIF.

2. Background
2.1 Possible language bias in large-scale assessments

The validity of comparisons of scores from large-scale assessments, such as PISA critically depends on whether the tests actually assess the same or similar knowledge and competencies across the groups and sub-groups that are compared (Hambleton, Merenda, & Spielberger, 2005). Most research concerning bias in large-scale, international assessments focuses on bias arising from incomparable constructs (tests measure different forms of knowledge and competencies) or measurement (the construct is the same but measurement is biased). In this study we investigate possible sources of bias for different language groups: students who attend schools with French as the language of instruction situated in English as the dominant language (Ontario) context and who speak or do not speak French at home. The latest PISA report suggests that in mathematics, “Canadian students continue to perform well in mathematics in a global context” (Brochu, Deussing, Home, & Chuy, 2012, p.18) scoring 24 points above the OECD mean, being outperformed by only nine countries. In Quebec, where many students take the test in French, the mean score was 18 points above the Canadian mean, whereas in Ontario, where most students take the test in English, the students achieved 4 points below the national mean. In fact, Quebec was the only province where students scored above the national mean; at the same time, in this province, there was the largest difference between low and high achievers.2 The difference between high and low achievers was lower than the national average in the province of Ontario. In

---

2 This difference is due to the larger proportion of high achievers; and it is observed in high-achieving countries generally (Brochu et al., 2012). The difference does not mean, however, that there is an issue with equity.
Canada, students in francophone school systems score higher than students in anglophone systems (Brochu et al., 2012).

What these PISA results do not provide are analyses of how the language spoken at home interacts with the language spoken at school with respect to PISA achievement, that is, the reported results do not tell us whether there is linguistic bias. This is an important topic in countries such as Canada where large parts of the population speak languages at home that differ from the school language. Linguistic bias may exist because of the incomparability and non-equivalence that exists not only between languages but also within languages (Derrida, 1996; Ricœur, 2004), such as when African American students are confronted with a different form of English at school, de facto making standard English a second language (Roth & Harama, 2000), or when Hawai’ian students do poorly when confronted with standard English literacy practices (Au, 1982). One key method for identifying bias exists in differential item functioning (DIF). It requires supplementation with reviews on the part of individuals with expertise in the relevant culture, cognition (here mathematical), language, or curriculum. This expertise allows the identification of the possible sources of differential item functioning embodied in the items. The identification of sources of DIF may also be conducted by means of think-aloud protocols in which individuals from the target populations think aloud as they solve the tasks of interest.

2.2. Differential item functioning

Differential item functioning (DIF) is one of the most commonly used methods for examining item comparability. An item exhibits DIF when it has different psychometric properties—e.g., item difficulty, discrimination—for students of equal ability in different groups. For example, for students from different French-speaking settings, the French terms taille [size], superficie [surface area], and aire [area]—all of which may be found in the same multilingual test (e.g., Roth et al., 2013)—may not be equivalent and therefore assess scientific and mathematical competencies related to the calculation of area in different ways. When different language groups are involved, DIF may arise from cultural references, linguistic (grammatical) complexity, familiarity/formality of words, differences in word use, pronouns, and others (Arffman, 2010; Roth et al., 2013). Such language-based differences are important given the fact that PISA items use contextual problems to elicit conceptual and procedural knowledge (Sáenz, 2009). DIF analyses establish the probability distribution \( f(Y|\Theta) \) of score \( Y \) given some latent trait \( \Theta \) (Osterlind & Everson, 2009). That is, DIF does not establish simple group differences but differences when examinees from the different groups compared are matched on the latent construct targeted by the assessment (Zumbo, 1999). Thus, the simple fact that students from Ontario attending French-language instruction schools have a significantly lower success rate on a PISA item does not in itself constitute DIF—e.g., see PISA item M124Q03 in Table 1 below—because the analysis accounts for the latent trait \( \Theta \) which turns out not to be the same. Rather, DIF arises when the group under consideration scores significantly above or below the scores predicted on the basis of the latent construct.
2.3. Expert analyses

To appropriately interpret the results from DIF analyses, the use of (external) expert reviews has been recommended across very different disciplines (e.g., Scott et al., 2010); expert reviews may detect sources of linguistic bias that have been overlooked during the official translation and testing of items (Arffman, 2010). Familiarity and knowledge with the context evoked in an item and through the particulars of the language used to express it are assumed to allow the expert to identify possible sources of differential item functioning. Because these sources may be of multiple nature, expert panels may be chosen such that the differences in their level of expertise might be used to differentiate curriculum issues—which a non-educator despite deep knowledge of language and culture might not detect—versus issues that have their sources in language alone, which an educator proficient but not expert in a language might detect (Roth et al., 2013). Although one study conducting blinded bilingual experts strongly recommended the use of experts (Scott et al., 2010), another study—which conducted think-aloud protocols of items with students—suggests that experts tended to focus on surface characteristics of items (Ercikan et al., 2010). The latter study stated that expert reviews alone could not be considered sufficient for the identification of sources of (linguistic) bias. This was confirmed in a more recent study of possible sources of DIF in a pan-Canadian science test administered in English and French. This study showed that experts with differing levels of expertise in language and curriculum did not reliably differentiate DIF and non-DIF items (Roth et al., 2013). The study had used think-aloud protocols to investigate the thinking of translation experts—some of whom had previously taught, others who had not—while assessing possible linguistic sources for biased test items.

2.4. Think-aloud protocols with examinees

Even if relevant content experts might be able to correctly identify sources of surface differences and DIF, they do not explain how specific item characteristics will lead to differential performances in the groups of interest. What examinees perceive in an item, what and how they are thinking, and where and why they make certain types of errors requires a research method that more or less directly assesses these processes. The think-aloud protocol is perhaps the most-used method for investigating what and how individuals—experts and novices alike—think while engaging problems and tasks of interest (Ericsson & Simon, 1993). It has been used to probe in great depth the reasoning scientists use when interpreting graphs (Roth & Bowen, 2001) or for the influences of language when English language learners solve mathematics word problems (Celedón-Pattichis, 2003; Goos & Galbraith, 1996). The fundamental assumptions underlying this method are that (a) individuals are able or can be taught quickly how to verbalize their thinking and (b) the verbalization process does not interfere with the task itself. To be consistent with this assumption, participants need to verbalize what is in their consciousness rather than reflecting about its processes and contents. The verbalizations then
constitute a protocol of what the individual has done to complete the task in the course of doing so. Recent research using functional magnetic resonance imaging (fMRI) while internal medicine physicians responded to multiple-choice questions provided supportive evidence—on the basis of statistically reliable differences in brain activities between answering and thinking-aloud—for the idea that the think-aloud protocol is a reasonable measure of thinking (Durning et al., 2013). Mathematics educators use the think-aloud protocol to investigate students’ and mathematicians’ mathematical thinking (including metacognitive strategies) and learning (e.g., Goos & Galbraith, 1996; Wilkerson-Jerde & Wilensky, 2011) or suggest using the think-aloud protocol for the investigation of students’ thinking during task engagement to supplement statistical analyses (e.g., Yang, 2012). Mathematics educators have exhibited the differences between thinking-aloud approach and clinical interviews, which exists for example in the fact that the prompts change the response and thinking patterns of the thinker (Koichu & Harel, 2007). These effects of the clinical interview exist even in the best cases, and, therefore, constitute limitations and constraints to this method.

3. Research design

This study was designed to better understand whether language might constitute one of the sources of differential item functioning in achievement tests when the students attend schools in a minority context and where students do or do not speak the school language at home. In the present study, we investigate possible language bias in nine published mathematics items from the 2000 and 2003 PISA administration—six of which are identified as exhibiting DIF when students attending French schools in Quebec were compared to students attending French schools in Ontario, where French was or was not the dominant language at home.

3.1. Study design

In this study, we employ a multi-level analysis of PISA mathematics items that may exhibit differential item functioning that has its source in language. Items were selected using item response theory-based DIF analysis (CTB/McGraw-Hill, 1991). Available items exhibiting DIF and some control items were subjected to three tiers of analysis. First, two bilingual university researchers with expertise in assessment and cognition reviewed the English and French version of the items. Then, a panel of four educators with expertise in the differences between French in majority, minority, and immigrant contexts assessed the French version of the items used in the present study. Thirty-three students attending a school in Ontario with French as instructional language who spoke at home either French or another language participated in think-aloud/interview sessions while solving the selected PISA mathematics items.

3.2. Participants, research protocols, and data sources collected

3.2.1. Review Panel 1: Educational research and mathematical cognition experts
A systematic language review of French DIF items was conducted, including comparisons with the English versions. The two fluently bilingual authors on our team conducted these analyses. One is a first-language French professor of education focusing on assessment of learning and large-scale assessment (Author3). The other is a professor of applied cognitive science focusing on the role of language in mathematics and science learning (Author1). His subject matter relevant background includes training as an applied mathematician, statistician, and mathematical physics; he taught science and mathematics for about 12 years. The reviews focused on (a) linguistic load of test items, (b) commonality of specific terms, (c) relevant cultural references, and (d) grammatical complexity. All comments by both panelists were summarized and discussed by the team. There were no notable disagreements.

3.2.2. Review Panel 2: Contextual language and curriculum experts

A panel of four French-speaking individuals with expertise in language and curriculum issues across the border of Quebec and Ontario reviewed all items for the purpose of assessing whether the items posed different challenges for students of different background, but all taking the test in French. Panelist 1 originally was from Quebec and had taught for a number of years prior to becoming a university professor specializing in assessment and evaluation generally and with a focus on the sciences in elementary and secondary particularly. Panelist 2 has 25 years of experience as teacher, school principal, and assessment specialist for the Ministry of Education, and currently is a curriculum principal for assessment in the province of Ontario, where she was raised and currently lives. Working closely with teachers, she has come to understand the complexity of assessment in practice. Her PhD focused on the trustworthiness of assessment. Panelist 3 is a university professor specializing in sociolinguistics and French as minority language originally from Ontario. She conducts research in French-language schools and was familiar with the PISA results and the relative achievement levels across Canadian provinces. Both Panelists 2 and 3 therefore represent speakers of French living in a minority context (Ontario). The fourth panelist is an immigrant, who has received her graduate degrees in Canada and who currently is a university professor. She is well versed in and has published on issues faced by immigrant students who speak a language at school that is not their mother tongue, especially in the subject area of the sciences (her undergraduate degree) and with a focus on learning in minority settings.

The panelists evaluated the items individually and then met in a group to discuss the issues that were arising from their analyses. Experts were provided with a kit to be used in the evaluation of the item stems and the specific questions associated with it. The evaluation grid categorizes item difficulty for semantic reasons (rare, academic, abstract, culture-specific, polysemic, familiar words not used at home, and false cognates), syntactical reasons (item length, number of words, multi-proposition sentences, passive voice, and verb tense), and other reasons (e.g., curriculum, structural incoherence, and other). There was space to justify opinions with examples and to provide open-ended, general comments. The panel was asked to indicate whether items possibly were biased for/against students attending
schools with French as instructional language who are (a) French native speakers from Quebec, (b) from French-speaking families in Ontario, or (c) from families speaking a language other than French; there was also an option to indicate whether no bias was anticipated. Panelists also were asked to indicate the degree of bias they expected. The completed kits were collected and entered in the database. The discussion following individual evaluation, led by one of the authors, was recorded and transcribed verbatim in its entirety. Similarities and differences between the responses are discussed in the presentations of the results. The verbatim comments were compiled and similarities and differences in their content were noted.

3.2.3. Students

The think-aloud part of the study was conducted at a private school in Ontario in which French is the language of instruction serving students from many different national backgrounds and with an equally diverse teaching staff. Although some studies suggest that the students attending private schools in Ontario do not differ from public school students (Roth, 2007), the possibility remains that the students from this particular school do not constitute a representative sample of Ontario students attending French-language instruction schools. Thirty-three students (18 female, 15 male) from the PISA target population (mean age = 15.6 years, SD = 0.2) were willing to participate. The students who did not speak French at home had attended French-language schools in Ontario for a mean of 6.1 years (SD = 4.2) whereas their peers who spoke French at home attended this type of school for 6.7 years (SD = 4.4). The languages other than French included Arabic (n = 4), Spanish (n = 2), English (n = 4), Romanian (n = 1), Flemish (n = 1), and Haitian Creole (n = 1), yielding a total of 13 OFNF and 20 OFF students in the present sample.

Students were invited to think aloud while doing the 15 PISA items, nine of which were from mathematics (6 reading items) following standard procedures for doing think aloud sessions. Each session lasted about 45 minutes. One or the other of the two authors conducting the sessions modeled thinking aloud followed by four practice items for the students to familiarize them with the protocol. This was deemed sufficient, as it had been suggested that training lasting a few minutes suffices to get participants to verbalize their thinking (van Someren, Barnard, & Sandberg, 1994). The students then solved the items while speaking aloud. When they fell silent, the respective researcher present reminded the student to think aloud by means of neutral statements (e.g., “Say what you think.”). After each item, the researchers also asked whether the student had encountered particular issues with the item, especially with the way in which it was presented. All sessions were audiotaped and subsequently transcribed verbatim. The student answer booklets were analyzed using the French scoring key provided together with the published items (OECD, 2011). All transcriptions were submitted to protocol analysis (Ericsson & Simon, 1993) to assess contents and processes of student reasoning while solving the PISA mathematics items. Because this part of the study focuses on how students think under test conditions, we distinguish between achievement and responses prior to probing and any changes therein that followed probing and questioning.
3.2.4. Participants for DIF analyses

The DIF analyses were conducted for the purpose of selecting candidate items in the three other types of analyses. Student samples for this study were identified from the PISA student background questionnaire item asking for the language spoken at home. We created three groups based on whether students spoke French at home or not and whether they lived in Ontario or Quebec: (a) Ontario students who attend school in French and speak French at home (OFF: \( n = 528, 2000; n = 284, 2003 \)), (b) Ontario students who attend school in French and do not speak French at home (OFNF: \( n = 427, 2000; n = 370, 2003 \)), and (c) Quebec students who speak French at home (QFF: \( n = 2,763, 2000; n = 1,846, 2003 \)). The results from the background questionnaire indicate that 5.6% of the OFF and 6.9% of the OFNF students were born outside of Canada. Because PISA uses a complex matrix sampling design, the total number of students responding to the (published) items investigated is much smaller than the total number of students from the populations who participated in the PISA 2000 and PISA 2003 assessments. This therefore limited the DIF analyses that could possibly be conducted.

3.3. Item selection

We conducted an analysis of all mathematics items from the 2000 and 2003 PISA administrations and identified all those that exhibited DIF in the comparisons of students attending French schools in majority settings (Quebec) and minority settings (Ontario), where students either speak French or another language at home.\(^3\) PISA explicitly states that items are stated in the everyday language of students rather than in mathematical terms (e.g., a mathematics register), which is to assure that students can relate each item to their everyday lives (OECD, 2003). For the DIF analyses, we distinguished between Quebec students attending French-instruction schools who speak French at home (QFF), Ontario students attending French-instruction schools who speak French at home (OFF), and Ontario students attending French-instruction schools who do not speak French at home (OFNF). The province of Ontario was chosen because the number of students required for conducting DIF analyses was too small in all other provinces. From the items that exhibited DIF, all those were included in our research that had been published together with their scoring rubrics. As a control, we also selected items that did not exhibit DIF and were associated with the same problem stem as the DIF items. This led to a total of six DIF and three non-DIF items: Walking (M124Q01, M124Q03), Continent Area (M148), Growing Up (M150Q01, M150Q03), Skateboard (M520Q01, M520Q02), Number Cubes (M555Q02), and Step Pattern (M806Q01). The success rates for the two populations of interest as well as those of the Quebec reference student population and the OECD means are provided in Table 1.\(^4\) PISA uses four

---

\(^3\) Simultaneously, we did the same for the published reading items exhibiting DIF but which are not investigated here.

\(^4\) The English items and their answer keys can be found in OECD (2009): Walking (pp. 101, 153), Continent Area (pp. 105, 158), Growing Up (pp. 106, 159–160), Skateboard (pp. 119, 169), Number Cubes (pp. 121, 170), and Step Pattern (pp. 124, 171).
item formats: multiple choice, complex multiple choice, closed constructed response, and open constructed response. In complex multiple choice items, students are required to choose more than one response. More complex response items receive partial credit marking and are scored polytomously (M124Q03, M148, M520Q01), whereas the remaining items are scored dichotomously (M124Q01, M150Q01, M150Q03, M520Q02, M555Q02, M806Q01). (The English version of the items used here are available in the supplementary materials.)

3.4. Analyses

To study the possibility of language bias of PISA items on student reasoning, we conducted DIF analyses to identify items on which students from different home language backgrounds attending French schools achieve differently. The DIF analyses were conducted using an item response theory-based approach where the item characteristics are compared taking the following statistical approach (Linn & Harnisch, 1981). The analyses were conducted using the Pardux software (CTB/McGraw-Hill, 1991) that implements this DIF detection procedure. In this approach, for each item the difference between observed and predicted mean response is computed for the groups of interest matched at different ability levels and for each decile of the item. The values for predicted probability of responding correctly \( P_{\text{pred}} \) were computed using the parameter estimates obtained from the entire sample and the \( \theta \) estimates for the members of the three groups considered here. The observed probability of responding correctly \( P_{\text{obs}} \) was obtained from the actual responses of examinees in the focal group data. The decision whether an item exhibits DIF is based on the evaluation of the \( p \) values of a Z statistic. Items are flagged as exhibiting DIF when \(|Z| > 2.58\) and the absolute value of the observed minus the expected \( p \) value is equal or greater than 0.10.

The results of the expert reviews of the items selected for the analysis (6 exhibiting DIF, 3 control not exhibiting DIF) were collated item by item. Summary statements were produced identifying all possible sources for DIF that the experts on the two panels identified collectively. Two researchers independently coded the data from the experts and negotiated any differences.

The think-aloud protocols were analyzed using standard methods (Ericsson & Simon, 1993). We followed other researchers who suggested taking the content of verbal protocols at face value and to assess these in their context (Ranking, 1988). For each item, the students’ response types and strategies for solving were identified, grouped, and counted. For those comparisons of the research that are in a theory-generation phase, we used Bayesian tests to establish the relative probabilities between null and alternative hypotheses (e.g., Wetzels et al., 2011). Standard statistical tests establish the probability of the data observed given the null hypothesis, \( p(\text{data}|H_0) \) and rejects the null when \( p \) is smaller than a critical value, generally \( p < .05 \). Bayesian tests, on the other hand, establish the probability of a hypothesis given the data, that is, \( p(H_0|\text{data}) \). When there are two hypotheses, \( H_0 \) and \( H_1 \), then \( p(H_1|\text{data}) = 1 - p(H_0|\text{data}) \). As a meta-analysis of 855 studies in experimental psychology shows, these tests prevent researchers from
overestimating the evidence in favor of effects. Given equal probabilities for two competing hypotheses prior to a study (i.e., \( p(H_0) = p(H_{alt}) = 0.5 \)), the JZS Bayes factor yields the relative posterior probabilities of \( p(H_0|D)/p(H_{alt}|D) \) given the data \( D \) at hand. The following classification scheme for the alternative hypothesis \( H_1 \) has been proposed: no evidence (\( JZS = 1 \)), anecdotal evidence (\( 1/3 \leq JZS < 1 \)), substantial evidence (\( 1/10 \leq JZS < 1/3 \)), strong evidence (\( 1/30 \leq JZS < 1/10 \)), very strong evidence (\( 1/100 \leq JZS < 1/30 \)), and extreme evidence (\( JZS < 1/100 \)) (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

4. Results

This study was designed to investigate linguistic and cognitive expert analyses and student think-aloud protocols of PISA items that may exhibit linguistic bias based on DIF analyses. The results of this study are summarized in five assertions:

1. The success rates observed within the student participant group tend to be higher than those of the population during PISA administration (Table 1 and Table 2);
2. The relative success rates within item groups could vary from those in the population (Table 1 and Table 2);
3. Expert analyses identified possible sources for linguistic bias even in items that DIF analyses suggested as not exhibiting language group differences;
4. Students often solved problems even though they encountered problems understanding certain words such that language did not appear to mediate the solution process; and
5. The possible sources of linguistic bias that experts identified did not affect actual performance as shown in the think-aloud protocol.

In the following, we exemplify these results by means of three representative analyses (Item sets Walking, Continent area, and Growing up). The items were selected because (a) all students had completed the items (see Table 2), (b) the selection represents items with negative, no, or positive bias based on DIF analyses (see Table 1), and (c) the results exhibited surprising differences to the trends in the population that may explain why population scores (sometimes) are low. For each of the 3 items (item sets), we present (a) items and results of DIF analyses, (b) analyses by research and cognition experts constituting Panel 1, (c) analyses by language in minority context and curriculum experts constituting Panel 2, (d) students’ solutions and analyses of think-aloud protocols, and (e) any evidence for differential item functioning based on language.

4.1. Walking (PISA 2000 items M124Q01, M124Q03)

4.1.1. Items and results of DIF analyses

5 A calculator for JZS values can be accessed at http://pcl.missouri.edu/.
The two items from this problem set provide an image of steps overlaid by a diagram indicating path length. The stem provides the formula \( \frac{n}{P} = 140 \) as relating number of steps per minute \((n)\) and step length \((P)\). The first of the two items asks students to calculate Henri’s path length given that he makes 70 steps per minute; the second item asks for Bernard’s speed (in m/min and km/h) given that his path length is 0.80 meters. The PISA result showed that the OECD mean for M124Q01 is 34% success rate (Table 1). The OECD mean for the second item (M124Q03) is 19%. Based on the criteria stated above, DIF analyses had revealed that M124Q01 was negatively biased against OFNF students as compared to QFF and negatively biased against the combined OFNF+OFF population; no DIF was identified for the QFF-OFF comparison. This suggests that a possible bias exists against those students who do not speak French at home. No DIF was detected for the second item M124Q03; it served as one of our control items.

4.1.2. Review Panel 1

The review of Panel 1 reveals that two of the words, consécutive [consecutive] and rapport [relation] may be difficult for students from Ontario. They also point to the three concepts involved traces de pas [footprints], longueurs de pas [pace length], and nombre de pas [number of steps] where the English uses clearly different terms. In addition, in the English version, the variable for pacelength starts with the letter \(P\), whereas this coincidence is not the case for the French. Students might take \(P\) to stand for the number of steps (pas). One panelist also suggested a possible confusion between pas [step] and the negation ne . . . pas [not], especially because in vernacular the ne of the ne . . . pas structure tends to be dropped.

4.1.3. Review Panel 2

Two panelists made note of sentences with multiple propositions and the length of individual propositions (e.g., la distance entre l’arrière de deux traces de pas consecutives [the distance between the rear of the two footprints]). Also noted was the possible confusion of pas with ne . . . pas. Three panelists took note of the academic nature of the text—e.g., traces de pas [footprints], consécutive [consecutive], ne . . . pas [not], and montrez vos calculs [show your work]—which might disadvantage students who do not speak French at home. A structural incoherence was noted in M124Q03, where the pacelength of Bernard is introduced prior to the statement that the formula provided also applies to him.

4.1.4. Student solutions and think-aloud protocol analyses

In the present sample, 45 percent of the students answered the first of the two items correctly \((SD = 0.51)\) (Table 2). There were no differences between students speaking French or another language at home. However, when the two groups were compared after controlling for the total score on the items from PISA 2000\(^6\), then there is substantial evidence \((JZS = 0.163)\) for the alternative hypothesis of

---

\(^6\) This is equivalent to controlling for ability when the total score is taken as a proxy measure thereof.
difference in favor of the OFNF students (i.e., the alternative is 6.13 times more likely than the null). In other words, the alternative hypothesis is 6.1 times more likely. (The bias in favor of OFNF compared to OFF remains when the missing solutions are coded 0 and the total score from all items is considered.) On item M124Q03, there is only anecdotal evidence in favor of the alternative hypothesis ($JZS = 0.637$) of students speaking French ($X_F = 1.19$, $SD_F = 1.17$) compared to those speaking another language at home ($X_{NF} = 0.58$, $SD_{NF} = 0.79$); that is, this alternative hypothesis is only 1.6 times as likely as the null hypothesis. The difference is even more accentuated given the fact that when the achievement on this item is residualized on their total score on PISA 2000 items, then there is a bias against students who do not speak French at home. The bias disappears when the missing solutions are coded 0 and the total score from all items is considered. Three students (9%) answered the item correctly, with 9 (27%) and 5 students (15%) receiving 2- and 1-point partial credit, respectively. Table 3 presents the number of students from each group and the solution (levels) they exhibited during the think-aloud sessions.

4.1.4.1. Item M124Q01

For the first of the two items, Table 1 shows that the success rate in the present sample was about the same as in the PISA 2000 assessment, with 7 (30%) and 5 (38%) of the students who speak French versus those who speak another language at home. The two most frequently used strategies leading to alternate solutions derived when students correctly substituted the number of steps but inappropriately solved for the unknown pacelength $P$ (e.g., either inverting the fraction involving 70 and 140 or multiplying the two numbers). One student did not substitute correctly, one student suggested that there were $2 \frac{1}{2}$ steps to be seen in the image and one student did not provide a response at all. The think-aloud protocols reveal that students frequently employed a *règle de trois* (rule of three) or *produit en croix* (cross product) in justifying their approach to solving the equation in the first question for $P$. The cross product students often placed a 1 below the 140 and then drew a cross or pair of arrows over the equation. Several students articulated an analogy such as $\frac{2}{1} = \frac{4}{2}$ or $\frac{3}{1} = \frac{6}{2}$ or $\frac{2}{4} = \frac{8}{4}$ before proceeding to solve the equation $\frac{70}{P} = 140$.

4.1.4.2. Item M124Q03

Tables 2 and 3 show that in the second of the two items, the frequency of students correctly solving for the unknown was higher than in the first item where the unknown had been in the denominator. The table also shows that all students receiving two partial credits ($n = 7$) or full credit ($n = 3$) were from the group speaking French at home. Three students in each group (i.e., 15% and 23%) did not attempt to solve the problem, frequently suggesting that it made no sense at all, that the item was fuzzy/vague (*flou*), or that there was something missing. Four students asked about what the 140 represented or why the relation between number of steps and pacelength was 140.
4.1.5. Evidence of (linguistic) bias from think-aloud protocols

Both expert groups had identified possible sources of linguistic bias. Two students, speaking Arabic and Flemish at home, verbally stated one number (60, 90) but wrote another number (70, 80). This facet arises from the particularity of French, where the numbers 70 (soixante-dix) and 90 (quatre-vingt-dix) are verbally expressed as 60 (soixante, 60) and 10 (dix) and 80 (quatre-vingt) and 10 (dix). Two students (LC016, LC031) queried how the equation can at all apply to the façon de marcher [way of walking]. Sevseveral students from both groups struggled with the complexity of the phrase la distance entre l’arrière de deux traces de pas consécutives [the distance between the rear of two consecutive footprints]. Three students asked about the noun rapport [relationship] and the adjective approximatif [approximate] (e.g., Ça veut dire quoi, “donne un rapport approximatif”? [That means what “gives an approximate relationship”?]). Three students either did not see a question at all or did not see a relation between the stem and item. One student (LC030) asked whether longueur de pas [pacelength] meant one step or two steps. Although the Romans, whose a pacelength, Lat. passum, used a double step in their measures, we do not have evidence that LC030 knows or does not know about this fact. Another student asked whether pacelength included the length of all steps of the foot (longueur de tout les pas du pied). Even though all of these students appeared to struggle with aspects of the stem and two items, they proceeded to work just with the formula and manipulated the numbers in one or another way.

4.2. Continent Area (PISA 2000 item M148)

4.2.1. Item and results of DIF analysis

The stem of this item shows a map of Antarctica together with a map scale. Students are asked to estimate the area of the continent using the map scale; they are to show and explain how they estimated the area. The DIF analyses revealed a bias in favor of Ontario students speaking French at home and the combined group of all Ontario students over their Quebec counterparts. The OECD mean for this item is 19%. Students receive full credit (2 points) if they use a correct approach and have a result between 12,000,000 and 18,000,000 km² and partial credit (1 point) for using a correct approach but estimating the area below or above.

4.2.2. Review Panel 1

The two experts did not identify potential sources for DIF based on language, and noted only that the term surface [surface] might be easier to understand than aire [limited portion of a surface, number measuring surface area], the term used in the Ontario curriculum guidelines.

4.2.3. Review Panel 2

It was noted that there was some fuzziness about the item, constituting a bias against all students, as the shadows used might confuse them about the actual area

---

7 In the English version, the text states less ambiguously “Bernard’s walking” rather than “way of walking.”
to be measured. A second possible confusion is between the scale, labeled Kilomètres, and the term échelle [scale] used in the stem. Another incongruence was identified to exist in using the term carte [map] with the map but to use the term échelle de cette carte [map scale] in the problem statement.

4.2.4. Student solutions and think-aloud protocol analyses

The statistical analyses of the PISA 2000 assessment show a bias in favor of Ontario students who do not speak French at home when compared to Quebec students; there is no such bias detectable when students speaking French at home are compared with Quebec students. In the present data, there are no statistically reliable differences ($p = 0.48$) between students speaking French ($M_{FR} = 0.85$, $SD_{FR} = 0.67$) and those speaking another language at home ($M_{OTH} = 0.77$, $SD_{OTH} = 0.83$) on this polytomously scored item. Six of the 33 students received full credit (2 points), corresponding to a success rate of 18% (Table 2), which is close to the OECD average of 19% (Table 1). Another 15 students (45%) received partial credit (1 point, Table 2) because they showed the correct approach but either over- or underestimated the dimensions of the continent or inappropriately scaled from the area of the map (in cm$^2$) to the surface area of the continent (Table 4).

Of the 24 students who calculated the area, one approached it as a circle, one as a diamond (two triangles), and 22 in terms of rectangles (squares) (Table 4). Those using the rectangle approach either calculated the map area of Antarctica in cm$^2$ and then scaled it ($n = 10$) or first measured the length(s) and width(s), scaled each to km, and then calculated the area in km$^2$ ($n = 12$). In the first instance, all students used the inappropriate scaling factor $F = \frac{1000(km)}{3(cm)}$. Those students using the second approach all ended up with an area within the range accepted as the standard response or over-/under-estimated one or both lengths and ended up with results somewhat below or above the accepted area. All of these results were of the correct order of magnitude. Five (15%) of the students actually measured the circumference in cm and then converted it to km, or used the scaling factor estimating how many times it fit into the circumference or calculated the circumference estimating the radius of a circle. Four students (12%) said that they knew nothing about area (aire) or that they are sure it had nothing to do with the circumference (périmètre).

Table 4 also provides evidence for the different perceptual experiences to which the task gave rise. Of those students who calculated the area, a large number perceived a rectangle or square ($n = 14$), a slanted rectangle ($n = 1$) or diamond ($n = 1$), or a circle ($n = 1$). Almost half as many, however, perceived the drawing in

---

8 The percentage of correct responses (18%) is deceiving as a measure of the cognitive competencies of these students, as some of the errors leading to partial or zero credit were of "clerical" nature. For example, one student verbally multiplies the estimated length and width generally (longueur fois largeur) and uses specific numbers (4,600 de largeur, 4,600 fois ça) and reports, after having used the calculator, reports and notes 9,600 km$^2$ as the result evidently having hit the “+” button.
terms of multiple rectangles involving the addition of never more than two smaller areas or, in one instance, a complex array of two larger areas from which smaller areas were subtracted that were included in the rectangle but not part of the continent.

LC005 said that she would have hoped to be asked for the perimeter, and then finds it to be 19,000 km, but then apologizes and states to have no idea about finding the area. LC012 first states not to know how to find the area when merely the scale is given and indicates that she does not well know the scale. She then constructs a polygon over the map, measures its extensions, finds the circumference and scales it to km. LC014 states that the task asks finding the area, which means “one has to like find, measure the outline (contour) with the ruler, and that’s it.” LC023 draws segments of 3 cm length around the outline of the continent and then, given the scale, multiplies by 1,000. LC030 states that the perimeter does not give the area, so that she does not know how to get the area.

Only two students made explicit reference to the reasonableness of the results they calculated. Thus, for example, LC033 divided the continent into two rectangles, measured the length and width of each, and scaled these to the size of the continent. He calculated the two areas and added them. He said that the result was large, and then noted, “because Antarctica is a continent, it would have to be large enough. And since 100 kilometers square is a surface area [superficie] that is large, I don’t know if it is too large, because I don’t have a sense for approximating” (LC033). The other student first provided a good estimate of the map area and then says that the scale is 2.9 cm to 1000 km but only divides the map area of 142.9 cm² by 2.9 to obtain 49.3 (km²) as his result. He then stated, “But Antarctica is a continent, is it really so small?” He then drew a square of 2.9 cm x 2.9 cm onto the map, gazed at it, and then stated an evaluation: “Yes I think this works. So this times 49. Two point 9 centimeters on the side, which equals one kilometer using the scale of the map, 1,000 km, I am sorry. I think this works.”

4.2.5. Evidence of (linguistic) bias from think-aloud protocols

Even though the DIF analyses flagged this item as showing positive bias for OFF and the combined OFF/OFNF group and even though experts identified issues that were supposed to lead to negative bias, there is little evidence in the think-aloud protocols for differential linguistic bias that would favor or disadvantage one or the other group. The only identifiable issue concerned the terms échelle [scale] and estimer [estimate]. Thus, LC002 (Arabic) was wondering about whether the word échelle (scale) refers to the scale next to the map, then suggests that the length is 1,000 km according to the scale, and then perhaps the area might be 1,000 times 1,000. LC019 (Romanian) wondered what the word estimate (estimer) might mean and then asks the researcher whether this is meant to say that one thinks. LC015 (French) read échelle as being used to say that “[the continent] is between 0 and 1,000” and “they want to know the area, like the area of a circle or a rectangle.” However, she could not do that because there were no measures given. She then asked what was meant by “estimating the area [C’est quoi, ça, qu’on demande d’estimer l’aire?].” Several students, even though they subsequently calculated the area, first stated not knowing what to do, because they had never learned to find an
area (aire) and then said that they had only dealt with (regular) forms, like the circle or square (e.g., LC024).

4.3. Growing up (PISA 2000 items M150Q01, M150Q03)

4.3.1. Items and results of DIF analyses

The stem of this item presents a graph with the average height of boys and girls in the Netherlands from 10 to 20 years of age (in the period from 11 to 13 years, the girls are taller than the boys) with the text describing this graph to represent the heights in 1998. The first of the two items asks what the average height of 20-year-old females was in 1980 given that the height has increased by 2.3 cm to 170.6 cm. Table 1 shows that QFF (71%) students had success rate 10% higher than the OECD (61%), whereas the OFF (57%) and OFNF (56%) students were slightly below the OECD. There was no DIF detected, and the item was included as a control. In item M150Q03, students are asked to explain how the graph shows that on average, the growth of the girls is slower [est plus lente] after 12 years of age. Although Quebec students (48%) were slightly more successful than the OECD average (46%), OFF (30%) and OFNF (23%) students were considerably below it. A cluster analysis for students from Nordic countries had shown that the two items belong to groups demanding different types of graphicacy: identification and critical analytical (Oleande, 2014). This would explain why the success rates on item M150Q03 (critical analytical cluster) ought to be lower than on M150Q03 (identification cluster). The DIF analyses show that OFNF students scored lower than what was expected in comparison with the QFF students, whereas there was no bias detected when OFF students were compared to QFF students. Because all Ontario students are exposed to the same curriculum, a potential source of bias OFNF is the language of the item.

4.3.2. Review Panel 1

The reviewers noted that in M150Q01, the second referent of depuis [since] is missing, as the stem refers to 1998 but the students responded to the item in 2000. The reviewers also noted the incongruence that the stem refers to taille moyenne des jeunes femmes [average height of young women], whereas the item M150Q03 text refers to filles [girls]. An issue with word familiarity was noted in that OFNF students would be more familiar with the word grandeur [height, Fr. grand = tall] than with taille [height]. The item also requires students to differentiate between taille [height] and croissance [growth], which, in the line graphs, are represented by the value of the curve at a point and the slope of the curve, respectively. Research showed that not only students fail to distinguish the values of the curve and its slope when asked to think aloud about unfamiliar graphs (Clement, 1989; McDermott, 1984) but also experienced scientists (Roth & Bowen, 2003; Roth & Middleton, 2006). A possible confusion arises because the referent for the comparison is not stated: Does the comparison est plus lente [is slower] take the boys’ graph as a referent or is the comparison within the girls’ graph but before and after 12 years?

---

9 The English version states “young females.”
Despite these possible sources for bias, the DIF analyses with the population data did not detect bias (Table 1).

4.3.3. Review Panel 2

In the version available to our research, the part of the stem *est représentée par le graphique ci-dessous* [is represented by the graph below] actually appears below the graph. Reviewers point out that this makes the sentence semantically incoherent. *Taille* is more frequently used to refer to the size and weight rather than height of a person. The panel took note that the question of item M150Q01 is not connected to the graph that so dominates the stem. Concerning item M150Q03, there was agreement of a potential difficulty that arises from the fact that in the graph, the population is described as *jeunes femmes* [young women], whereas the question refers to *filles* [girls]. Several reviewers noted a potential difficulty that arises from the less familiar phrase *en quoi* [in what respect]. Finally, in the same way as panel 1, panel 2 participants suggested that students may not understand the referent of the comparison. Although the panel identified possible sources, consistent with the significantly lower scores of OFF and OFNF when compared to QFF (see Table 1), there was no evidence for DIF.

4.3.4. Student solutions and think-aloud protocol analyses

4.3.4.1. Item M150Q1

On item M150Q01, the success rates of students from the population speaking French (71%) versus those who do not speak French at home (75%) (Table 2) are close to those of QFF students (71%) and about the OECD rate (61%) (Table 1). The null hypothesis is 2.9 times as likely as the alternative (*JZS* = 2.92). On item M150Q03, the success rates were 81% (OFF) and 83% (OFNF) (Table 2), respectively, considerably above the reported OECD rate (46%) and those of the population (Table 1). The null hypothesis is 3.0 times more likely than the alternative (*JZS* = 3.00).

The correct answer was frequently stated very quickly in general terms. More interesting and revealing, perhaps, are the responses and solution paths of those students scored unsuccessful using the answer key. Of these, one (3%) was a simple calculation rather than conceptual error (170.6 – 2.3 = 167.9). Three students (9%) said that there was not enough information or that the information pertained to two different years. Two of these students noted that the graph provides information only for 1998. That is, this is precisely the problem that was identified by the experts on panel 2. Four students (12%) said that the girls were 170.6 centimeters tall, reasoning as exemplified in the following example:

They tell me that the average height has increased, and then they give me the answer. . . . This is not clear, like if 170.6 is the, it’s what like, if it is the average of the young girls of 20 years, and then they ask me the same question. [LC012]

In all of such cases, the students read the question *as if* it said that by 1980, the height of the young girls had increased by 2.3 cm so that the answer was already provided. LC026 concludes that the question implies the height in 1980 to be different from that in 1998. He then suggested having to look at the graph and
states, “This graph is a graph from 1998, and they are asking about 20 years. I think, so I am deducing therefore that they must have been around 170 centimeters, because 18 years before . . .” and then did not continue.

Thirteen students looked at, referred to, or said having to go to the graph. Those 7 with successful solutions then tended to state that they only had to subtract the numbers given in the question (e.g., “No, it suffices to take away 2.3 from 170.6”). LC030 also refers to the graph, from which he takes that the average height in 1998 is 172 centimeters. As Figure 1 shows, the estimate of 172 centimeters is very close. Because the student calculated the height in 1980 to have been 172 – 2.3 = 169.7 cm, the incorrect solution has arisen from the incongruence between verbally and visually presented information. LC004 also identified the height as 172 from the graph, but then took the height given in the text (170.6 cm) to calculate the response. LC009 and LC015 found it bizarre that the graph for 1980 was omitted and, therefore, that a comparison of graphs could not be conducted. So they simply subtracted the number in the question statement. For the same reason, LC011 and LC023 said that they could not respond to the question. Notably, one of the members of Expert Panel 2 also felt that there were data missing, wondering about the exact period that the 2.3 cm increase had happened (“I noted the incoherence in the data. Students are asked to make statements when they do not even have the necessary data to respond to the question” [J4]).

4.3.4.2. Item M150Q3

A total of 29 students completed Item M150Q03 correctly (Table 2). It was successfully solved by 90% (n = 19) of the Ontario students speaking French at home and by 77% (n = 10) of their peers from families that speak a language other than French at home. These means considerably exceed those of the OECD, QFF, OFF, and OFNF populations (Table 1). Moreover, whereas in the comparison of the OFF and OFNF populations, the latter was less successful, the data here provide anecdotal evidence in favor of the null hypothesis (JZS = 2.41).

Most of the students in the combined group (n = 20, 61%) explained the way in which the graph shows that the girls’ growth slows after the age of 12 by using qualitative, everyday language. Students used expressions such as augmente fortement/faiblement/plus lentement [increases strongly, weakly, more slowly], monte très lentement/rapidement [increases very slowly/quickly], se bloque [is stuck], stagné [stagnates], évolution rapide [quick evolution], pique [stings], s’aplatit [flattens], tendance à arrêter de grandir [tendency to stop growing], and reste droite [stays even] to describe what the curve does. Two students (6%) provide explanations using mathematical language, both using the term pente [slope]. Seven students (21%) use numbers to compare the growth between 10 and 12 years to that for the same or longer periods after 12 years. Even students who were less successful overall on the set of items provided acceptable answers to this item. Thus, for example, LC026, who together with one other student had the lowest overall score, said:

You can see that after 12 years, the growth is slower, because the curve is slightly less increasing than between 10 and 12 years, where the growth is
steep. It stabilizes approximately after attaining 12 years, in contrast to the men. (LC026)

Four students did not provide responses that receive credit according to the answer key (Table 2). One student said, "I don’t see any element that could prove to me that it is slower, that it is slower what?" (LC005). Two students [LC022, LC025] compared the lines of boys and girls at or some time after the age of 12—e.g., la ligne monte plus vite après 13 ans que les filles [the line increases much more rapidly after 13 years than the girls] (LC022—as if they had read the comparative est plus lente [is much slower] in reference to the curve of the boys rather than as an internal comparison with the curve for the girls prior to 12, that is, as if they had read “slower . . . than that of the boys.” Finally, one student made reference only to the growth of women at 20.

4.3.5. Evidence for (linguistic) bias from think-aloud protocols

The analyses of those 7 think-aloud protocols where students did not obtain the target response because there was not enough information or because the answer given reveals that the source of the problem does not lie in mathematical skill but in the way in which students read and understood the text. There is evidence that for a number of students the temporal relations are at issue (and perhaps reference of the depuis). Thus, LC002 read the text in conditional form (en 1980 . . . pourra atteindre [in 1980 . . . could reach]; LC008 and LC012 read the text as if it stated that by 1980, the height had increased to reach 170.6 cm; and LC011 asked, depuis 1980 . . . pour atteindre . . . en? [since 1980 . . . to reach . . . in?]. LC020 stated not quite well understanding the relationship between en 1980 [in 1980] and ici, de 2.3 centimètres pour atteindre 170.6 centimètres [here, by 2.3 cm to attain 170.6 cm]. LC029 and LC031 explicitly noted that the question did not state the date of reference (“do not say which date it is”; “They don’t tell me when . . .”) There may be more students who might have read the question differently but may have simply resorted to subtracting the numbers because of the relation has increased by . . . to (LC029, “They are giving us the data I do know from when, they are not saying, but because they are saying has increased by 2.3 centimeters and give the value after the increase, I just do a subtraction”). LC034 noted that the question has no relation to the graph shown, and then explains that only a subtraction is required. LC025: "So we are in 2000, but then they are asking about 1980"; LC022, “They could have perhaps stated until today . . .” LC012, “Since 1980, it has increased by 2.3 cm, so now [maintenant] . . . I don’t quite understand the question” LC010, “They are not giving me the year, they only give me 1998”

To investigate whether there is a possible source of differential item functioning in item M150Q03, we compared the relative success rates in the different answer categories. Although 33% of the OFNF students (n = 4) compared to 14% (n = 3) of the OFF students chose the mathematical approach, the difference does not reach statistical significance ($\chi^2 = 1.66, p > .05$). By contrast, 2 OFF students (10%) used a mathematical term, whereas no OFNF student used mathematical terms.\(^{10}\) Exactly half of the OFNF students (n = 6) used vernacular language, whereas 66% (n = 14)

---

\(^{10}\) Empty cells or small numbers in cells prohibit $\chi^2$ tests.
of the OFF students did so ($\chi^2 = 0.88, p > .05$). OFNF students provided both of the non-valid items making boy/girl comparisons (16%).

5. Discussion and implications

This multi-method study was designed to investigate the role that language might play in differential item functioning of an international test when administered to students living in a minority situation (i.e., language of instruction is not the dominant language of the embedding culture) and who do not speak the test language (i.e., language of instruction) at home. The results reveal both confirmation and some surprising differences on the PISA assessment. For example, in the open constructed response item M150Q03 (Growing up), which might be considered to draw heavily on language expertise, the rate of success among OFF and OFNF students was unexpectedly high. In the think-aloud situation, the students may express themselves much more so than in a one-sentence response in the paper-and-pencil format. That is, the fact that they conceptually understand what is at stake is communicated, because the format allows this to happen. The present study thereby reveals some possible reasons for students’ lower success rates under paper-and-pencil conditions versus the think-aloud protocols: because students could articulate themselves at length and in their vernacular about the trends in graphs, they did in fact produce answers that were correct according to the PISA coding scheme, which they might not have been able to do if they were producing a one-sentence written response. The data and our analyses led us to think that a fruitful hypothesis to pursue in future research might be that PISA may not assess conceptual understanding and mathematical literacy. Instead, PISA might allow students to be successful even when they do not understand because of the language, because (a) they are familiar with the background (e.g., skateboarding), know specific (technical) words or (b) the grammatical structure renders its sense ambiguous or inhibits comprehension.

The think-aloud protocols show that even when there were issues with the language, students often correctly solved the item. That is, although the experts on both panels might have identified items that could lead to differential item functioning, this did not appear to happen. As reported in another study focusing on students in language minority settings, when students do not know or understand a word, they simply continue (Celedón-Pattichis, 2003). In contrast to Celedón-Pattichis’ study, however, our students generally did not (overtly) translate; and there are only two instances where implicit translations—skateboard terms understood in terms of their English equivalents and the use of dollars and Euros instead of “Zeds.” The examples this author provides show, however, that her students were much less fluent than the students in the present study were and yet, even in that case the conclusion was that “language did not have a significant effect on word problems” (p. 87). Our think-aloud protocols provide an answer for this apparent riddle: If a formula was given, students used it (“If I did not understand the text, I would simply use the formula, because it is the only thing that gives me a reasoning,” LC032). If the item led students to think that something had to be added (as in Skates), the students added (sometimes too little or too much). Thus,
independent of the textual particulars, students have learned to respond in some way that the item structure seems to suggest. One student explicitly noted: “Even if I read the text and I didn’t understand, all I do is take this formula” (LC032).

Mathematics educators have discussed the role of context in contextual word problems; any context essentially is created by verbal descriptions (e.g., Roth, 1996). This might be used to hypothesize that language constitutes an essential mediating factor of achievement. This hypothesis would be supported by the lower achievement means of OFF and OFNF when compared to the QFF students (Table 1), and, therefore, point to the direct effect of language on group differences in mathematical achievement that another think-aloud study of mathematics items reports (Ercikan et al., 2010). In the present study, the data provide little support for the hypothesis differences between students who speak versus do not speak the instructional language at home. But does language also lead to differential item functioning? The results show that in our sample, when overall achievement was controlled, there was little evidence for differential effects. That is, across the items exhibiting DIF in the population, the data collected here provides at best anecdotal support for differential item function. If the DIF at the population level is a real effect, then our sample may actually represent a sub-population in which the overall trends do not hold. The identification of such sub-population would be an important area of future investigation (Ercikan & Roth, 2014).

There have been suggestions that language plays an important role of mathematical understanding and sense-making, and students who study the subject in a language other than their mother tongue face special and a diversity of challenges that make it difficult to make generalizations about their instructional needs (Moschkovich, 2011). Although PISA items are constructed to make sense, our participants frequently did not think so. Several students complained about the seemingly arbitrary nature of the situation or the way in which it is stated. They used words such as “bizarre” \( (n = 10) \), “ambiguous,” “fuzzy \([flou]\),” “just about anything \([n’importe quoi]\),” “not realistic,” and “absurd and arbitrary” to describe what they were confronted with. It is notable that in the present instance, where the text makes reference to a very mundane experience, walking, only three students (LC021, LC028, LC033, LC034) made reference to the relation between the results of their calculations and the reasonableness in their world—e.g., whether it is reasonable to have a pacelength of 0.5, 2, 20, or 9,800 (m), or whether the calculated speed, in m/min or km/h is reasonable given their experience of walking. For the other students, the real problem was one of replacing the letters in the equation by a number (“because this is what they are asking us to do” LC025). There was no relation between some everyday world situation and what the students were doing, producing, and understanding to have done. There was one student (LC017) who said that 9,800 makes no sense and then changed from 140 x 70 to 140/70; he also arrived at calculating 175 steps per minute, which he said makes no sense because he wants meters per minute. Two students (LC033, LC034) had calculated \( P = 140/70 \), then said that 2 meters made no sense for the pacelength, and then calculated \( 70/140 = \frac{1}{2} = 0.5 \text{ m} \). Several students “complained” about the fact that there were no explanations as to the nature of the rapport [relation] \( n/P = 140 \). One student in particular talked about the supposedly realistic nature of word problems,
but that sometimes teachers put numbers into problems that have nothing to do
with reality: “Sometimes teachers give exercises that have nothing to do, and one
finds, for example, a person who buys 420 liters of milk because his cow…”
(LC028); and the student continues commenting, “sometimes these things are
completely wacky, and so I have no means to know whether its good or not.”

This study also shows that similar to the experts, students sometimes offered
alternative ways of stating a problem that would avoid structural problems. Thus,
for example, LC005 offered an alternative framing, Quel est le prix maximum et
minimum en assemblant les éléments soi-même, quelque chose comme ça [What is the
minimum and maximum price when the elements themselves are assembled, or
something like that], using in assemblant [assembled] precisely the same verb that
experts had proposed.

Differential item functioning, as pointed out above, does not reflect the
differences between group means. Instead, it is the differential effect arising from
comparing expected with actual achievement distribution of an item given a latent
trait (here ability as measured by the overall score). If a group is expected to score
below another one based on the overall mean, and it achieves as expected, then
there is no DIF. Instead, DIF exists when the group of interest achieves significantly
below or above the expected level. This may allow us to understand why the experts
tend to point to those linguistic issues that lead to differences between group
means. These are only surface characteristics from the DIF perspective.

Simultaneously, experts fail to identify the much more difficult secondary
differences of DIF, where achievement may be higher than expected (positive bias)
but still considerably below the comparison group. Thus, for example, in this study
the OFF group alone and combined with OFNF had a lower mean on item M148 than
the QFF population and yet there was a positive bias for the former. Future studies
may investigate whether experts can tease out the differential rather than the
overall effects of language on performance differences.

This study advances our discipline’s understanding of possible sources for
differential item functioning in large-scale tests of mathematical literacy. The nature
of this study, which used multiple methods not normally found together in one
investigation, affords identifying possible problems in single-method studies. On the
one hand, there are studies that use experts only (e.g., Arffman, 2010; Roth et al.,
2013). The present study shows that experts tended to highlight issues that might
be expected to lead to the lower performances of the two populations of interest
(OFF, OFNF) compared to those of French speakers living in a majority context
(Québec) and OECD means. Although the think-aloud protocols did not exhibit lower
performance based on students’ linguistic fluency, especially when they did not
speak French at home, the fact that a proportionately higher number of OFNF did
not get to complete the items may be an indication of the observed lower average
achievement of students in minority settings. Language and education experts
appearing in other studies pointed out slower reading and higher demands on
cognitive processing as possible sources for lower achievement (Roth et al., 2013).
Because the students in this study all attended a private school, there may be a
selection bias so that the students constitute a sub-population where the differences
observed in the general student population in the geographical area (Ontario) do
not exist. However, other studies appear to show that selecting private school students in Ontario does not constitute selection bias (Roth, 2007). Further study is required selecting students from public schools to investigate whether these light to changes in the probabilities of the hypotheses reported here. Moreover, future studies may also attempt to capture the extent to which students speak in the school, home, or another language when they are with friends in out-of-school contexts.

At the same time, the identification of item bias appears to be more complex than we anticipated. The experts appear to focus on direct effects of language mathematical performance rather than the differential effects. This is so even though both panels knew or were made aware of the fact that the differential effects were of interest to the study rather than simple group differences. Further research is required to understand how think-aloud protocols, conducted at the individual level, can be used to reveal effects that have been identified by DIF analyses at the group level. For example, possible sources of DIF were discussed on expert panel 2. Panelists suggested as a possible hypothesis for items that exhibit bias in favor of students who do not speak French at home: Those families may in fact be of higher socio-economic status and cultural milieu interested in their children learning to be fluent in French even though their dominant language is English (or some other language). A panelist familiar with French language in minority situations cited a study that found school boards to serve such students better than their regular students. Moreover, these students then might be finding other strategies for coping with situations in which they are confronted with unfamiliar language—as shown, for example, in those participating students who saw the skateboard item in terms of the English equivalent terms or students who, without further reflecting, talked/thought about the prices given in terms of dollars or Euros. This could possibly explain the differential effect of the items, which, on the one hand would bias against OFNF students because their fluency is less than that of other students taking the test in French; and, on the other hand, these student may do better than expected because they are from backgrounds that normally provide a positive bias. This might be particularly the case in—and actually a confounding thus limiting factor of—the present study conducted in a private school with tuition fees that would make it inaccessible for some (though there is free admission in special cases). On the other hand, students from working class families speaking a language other than French at home may seek attending French-language schools at a (much) lower frequency than students from more affluent homes. Future research is necessary to investigate the relationship between fluency, socio-economic status, achievement, and forms of reasoning as revealed in think-aloud protocols.

Contrary to the assumption that think-aloud protocols do not interfere with thinking (Ericsson & Simon, 1993), there were repeated instances in our database where students suggested not being able to think and verbalize their thoughts. Future research is required to tease apart how this multitude of factors might influence results of PISA and similar assessments. Our research group currently considers investigating the similarities and differences between investigating student reasoning by means of the think-aloud protocol versus retrospective accounts, possibly stimulated using videotapes of the test situation.
Psychometricians have suggested that results true for a population may in fact not apply at all to individuals or sub-groups of the population (Boorsboom, Mellenbergh, & van Heerden, 2003; Ercikan & Roth, 2014). This is so because experimental and survey research designs and related statistical analyses, are based on inter-individual differences whereas the psychological effects of a treatment or actual problem solving behavior are based on intra-individual differences. Similarly DIF analyses focus on inter-individual differences and compare response patterns at the group levels. Previous research has provided evidence that DIF findings at the group level may not hold for sub-groups (Ercikan & Oliveri, 2014; Ercikan et al., 2014; Oliveri, 2012). Observations at the individual level, such as think-aloud protocols, may not reflect group- and population-level trends, yet may be able to provide insights about generalizability of findings at the group level to individuals and sub-groups. Future research may be designed to collect information for identifying DIF and think-aloud information simultaneously from the same individuals, which allows locating them in the intra- and inter-individual paradigms.

6. Coda

This study was designed to investigate language as a possible factor in producing differential item functioning. Our study shows that students attending French-instruction schools do not constitute a homogeneous group, as those students attending school in Quebec outperform those who attend school in Ontario. Among the latter, those students who speak French at home outperform those who do not speak French at home. Yet this pattern is complicated by the fact that there are differential effects such that one or the other group of students performs below or above the overall trend. Our study also showed that the general overall trends and the differential effects may not hold for students attending private schools, or private schools such as the one participating in our research. Although this study focused on sub-populations of students attending French-instruction schools in Canada, similar investigations might be conducted in other countries wherever the home language is different from the school language, even if the two are versions of the same language, such as in the case of standard and non-standard English (Roth & Haraama, 2000), high German and Swiss German, or Catalan and Spanish. Finally, linguistic, psychometric, and cognitive experts tended to identify language issues that contribute to group differences rather than to differential item functioning, a result particularly salient with those items where linguistic bias was positive for the disadvantaged group.

Acknowledgments

This study was supported by grant #410-2011-0745 from the Social Sciences and Humanities Research Council of Canada.

References


**Captions**

Fig. 1. There is a contradiction between the height of 20-year old women as stated in the text and in the actual value of the graph.
Table 1. Success rates on selected PISA items, six of which exhibit differential item functioning, in the two target populations, the Quebec reference population, and reported OECD means

<table>
<thead>
<tr>
<th>Item</th>
<th>DIF QFF compared to (^1)</th>
<th>OECD</th>
<th>Success rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>QFF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OFF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OFNF</td>
</tr>
<tr>
<td>Walking (M124Q01)</td>
<td>-OFNF, -(OFF, OFNF)(^2)</td>
<td>34</td>
<td>51</td>
</tr>
<tr>
<td>Walking (M124Q03)</td>
<td></td>
<td>19</td>
<td>55</td>
</tr>
<tr>
<td>Continent area (M148)</td>
<td>+OFF, +(OFF, OFNF)</td>
<td>19</td>
<td>45</td>
</tr>
<tr>
<td>Growing up (M150Q01)</td>
<td></td>
<td>61</td>
<td>71</td>
</tr>
<tr>
<td>Growing up (M150Q03)</td>
<td>-OFNF</td>
<td>46</td>
<td>48</td>
</tr>
<tr>
<td>Skateboard (M520Q01)</td>
<td></td>
<td>72</td>
<td>83</td>
</tr>
<tr>
<td>Skateboard (M520Q02)</td>
<td>-(OFF, OFNF)</td>
<td>46</td>
<td>67</td>
</tr>
<tr>
<td>Number cubes (M555Q02)</td>
<td>+(OFF, OFNF)</td>
<td>63</td>
<td>70</td>
</tr>
<tr>
<td>Step pattern (M806Q01)</td>
<td>+OFF, +(OFF, OFNF)</td>
<td>66</td>
<td>62</td>
</tr>
</tbody>
</table>

Note 1: Ontario students attending French instruction schools and speaking French (OFF) or another language at home (OFNF) at home (minority setting) are compared with students from Quebec attending French instruction schools and speaking French at home (QFF) (majority setting). The “+” and “-” signs indicate DIF for and against the Ontario students.

Note 2: Parentheses indicate that the bias is against the combined score of OFF and OFNF.
<table>
<thead>
<tr>
<th>Item</th>
<th>Total points</th>
<th>INC¹</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking (M124Q01)</td>
<td>1</td>
<td>0</td>
<td>18</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walking (M124Q03)</td>
<td>3</td>
<td>0</td>
<td>16</td>
<td>5</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Continent area (M148)</td>
<td>2</td>
<td>0</td>
<td>12</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Growing up (M150Q01)</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Growing up (M150Q03)</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skateboard (M520Q01)</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skateboard (M520Q02)</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number cubes (M555Q02)</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step pattern (M806Q01)</td>
<td>1</td>
<td>10</td>
<td>4</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹ Number of students who did not attempt item
Table 3. Results for the Walking (M124Q01, M124Q03) PISA item for students attending a French-language instruction school in a minority setting speaking French ("French") versus students speaking another language at home ("Other").

<table>
<thead>
<tr>
<th>Item</th>
<th>Solution</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking Q1</td>
<td>Full credit</td>
<td>Other (n = 13)</td>
</tr>
<tr>
<td></td>
<td>1. ( \frac{n}{P} = 140 \rightarrow P = \frac{n}{140} = 70 ) = 0.5</td>
<td>5 (38%)</td>
</tr>
<tr>
<td></td>
<td>2. ( \frac{70}{P} = 140 \rightarrow P = \frac{70}{140} = \frac{1}{2} ) = 0.5m</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. ( \frac{70}{P} = 140 \rightarrow P = \frac{140}{70} = 2 ) = 20m</td>
<td>6 (46%)</td>
</tr>
<tr>
<td></td>
<td>2. ( \frac{70}{P} = 140 \rightarrow P = 140 \times 70 = 9800 )</td>
<td>1 (8%)</td>
</tr>
<tr>
<td></td>
<td>3. Other (( \frac{70}{P} = 14 ) and 2 ½)</td>
<td>1 (15%)</td>
</tr>
<tr>
<td></td>
<td>4. none</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Walking Q3</td>
<td>Partial credit (1)</td>
<td>7 (54%)</td>
</tr>
<tr>
<td></td>
<td>( \frac{n}{P} = 140 ), ( \frac{n}{0.8} = 140 \times 0.8 = 112 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. 112*0.8 = 89.6</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>2. ( v(km/h) = 112 \times 60/1000 = 6.72 )</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>Full credit (3)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td></td>
<td>112*0.8 = 89.6, ( v(km/h) = 89.6 \times 60/1000 = 5.376km/h )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. ( \frac{n}{P} = 140 \times 0.8 = 7952 )</td>
<td>3 (23%)</td>
</tr>
<tr>
<td></td>
<td>2. 0.8*60 = 48(steps/m)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. ( \frac{n}{P} = 140 \times 0.8 = 140 \times 0.8 ) = 175 \times 175* 0.8 = 140(m/min) = 14(km/min)</td>
<td>3 (23%)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 (23%)</td>
</tr>
</tbody>
</table>
Table 4. Results for the Continent Area (M148Q02) PISA item for students attending an Ontario school with French as instructional language

<table>
<thead>
<tr>
<th>Approach</th>
<th>N (33)</th>
<th>OECD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Length x width) Map (cm²) → conversion</td>
<td>10</td>
<td>(30%)</td>
</tr>
<tr>
<td>Conversion (km) → (Length x width) (km²)</td>
<td>12</td>
<td>(36%)</td>
</tr>
<tr>
<td>Diamond</td>
<td>1</td>
<td>(3%)</td>
</tr>
<tr>
<td>Circle</td>
<td>1</td>
<td>(3%)</td>
</tr>
<tr>
<td>Perimeter</td>
<td>5</td>
<td>(15%)</td>
</tr>
<tr>
<td>None</td>
<td>4</td>
<td>(12%)</td>
</tr>
<tr>
<td>One area</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rectangle, square:</td>
<td>14</td>
<td>(42%)</td>
</tr>
<tr>
<td>Circle:</td>
<td>1</td>
<td>(3%)</td>
</tr>
<tr>
<td>Rectangle on side, diamond:</td>
<td>2</td>
<td>(6%)</td>
</tr>
<tr>
<td>Multiple areas, rectangles (added, subtracted):</td>
<td>7</td>
<td>(21%)</td>
</tr>
</tbody>
</table>
Released PISA Items Used in this Study

The items may be found in OECD, 2009, pp. 101, 105, 106, 119, 121, and 124.

The picture shows the footprints of a man walking. The pacelength $P$ is the distance between the rear of two consecutive footprints.

For men, the formula, $\frac{n}{P} = 140$, gives an approximate relationship between $n$ and $P$ where,

$n = \text{number of steps per minute, and}$

$P = \text{pace length in metres.}$

**QUESTION 2.1**

If the formula applies to Heiko’s walking and Heiko takes 70 steps per minute, what is Heiko’s pacelength? Show your work.

**QUESTION 2.2**

Bernard knows his pacelength is 0.80 metres. The formula applies to Bernard’s walking.

Calculate Bernard’s walking speed in metres per minute and in kilometres per hour. Show your working out.
QUESTION 5.1

Estimate the area of Antarctica using the map scale.

Show your working out and explain how you made your estimate. (You can draw over the map if it helps you with your estimation)
MATHEMATICS UNIT 6: GROWING UP

Youth grows taller
In 1998 the average height of both young males and young females in the Netherlands is represented in this graph.

QUESTION 6.1
Since 1980 the average height of 20-year-old females has increased by 2.3 cm, to 170.6 cm. What was the average height of a 20-year-old female in 1980?
Answer: ____________ cm

QUESTION 6.3
According to this graph, on average, during which period in their life are females taller than males of the same age?
MATHEMATICS UNIT 21: SKATEBOARD

Eric is a great skateboard fan. He visits a shop named SKATERS to check some prices.

At this shop you can buy a complete board. Or you can buy a deck, a set of 4 wheels, a set of 2 trucks and a set of hardware, and assemble your own board.

The prices for the shop’s products are:

<table>
<thead>
<tr>
<th>Product</th>
<th>Price in zeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete skateboard</td>
<td>82 or 84</td>
</tr>
<tr>
<td>Deck</td>
<td>40, 60 or 65</td>
</tr>
<tr>
<td>One set of 4 Wheels</td>
<td>14 or 36</td>
</tr>
<tr>
<td>One set of 2 Trucks</td>
<td>16</td>
</tr>
<tr>
<td>One set of hardware (bearings, rubber pads, bolts and nuts)</td>
<td>10 or 20</td>
</tr>
</tbody>
</table>

**QUESTION 21.1**

Eric wants to assemble his own skateboard. What is the minimum price and the maximum price in this shop for self-assembled skateboards?

(a) Minimum price: ............ zeds.
(b) Maximum price: ............ zeds.

**QUESTION 21.2**

The shop offers three different decks, two different sets of wheels and two different sets of hardware. There is only one choice for a set of trucks.

How many different skateboards can Eric construct?

A. 6  
B. 8  
C. 10  
D. 12
MATHEMATICS UNIT 23: NUMBER CUBES

QUESTION 23.1
On the right, there is a picture of two dice.

Dice are special number cubes for which the following rule applies:

The total number of dots on two opposite faces is always seven.

You can make a simple number cube by cutting, folding and gluing cardboard. This can be done in many ways. In the figure below you can see four cuttings that can be used to make cubes, with dots on the sides.

Which of the following shapes can be folded together to form a cube that obeys the rule that the sum of opposite faces is 7? For each shape, circle either “Yes” or “No” in the table below.

<table>
<thead>
<tr>
<th>Shape</th>
<th>Obeys the rule that the sum of opposite faces is 7?</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td></td>
</tr>
</tbody>
</table>
QUESTION 26.1
Robert builds a step pattern using squares. Here are the stages he follows.

As you can see, he uses one square for Stage 1, three squares for Stage 2 and six for Stage 3.

How many squares should he use for the fourth stage?

Answer: ................................. squares.