frontiers in Psychology

ORIGINAL RESEARCH
published: 27 July 2021
doi: 10.3389/fpsyg.2021.657335

Differential Item Functioning Among English Language Learners on a Large-Scale Mathematics Assessment

Ruixue Liu[1]* *and Kelly D. Bradley*[2]

[1] Department of Psychology, Henan Normal University, Xinxiang, China, [2] Department of Educational Policy Studies and Evaluation, University of Kentucky, Lexington, KY, United States

The current study research showed the nature and potential sources of the gaps in mathematics achievement between English language learners (ELLs) and non-English language learners (non-ELLs). The nature of achievement gap was examined using three DIF methodologies: including Mantel-Haenszel procedure, Rasch model, and Hierarchical Generalized Linear Model (HGLM). These were conducted at the item level in contrast to total test level. Results revealed that the three DIF approaches identified 10 common items. These 10 items demonstrated in favor of non-ELLs. Findings from this study will help educational researchers, administrators, and policymakers understand the nature of the achievement gap in mathematics at item level so that United States can be competitive in middle school mathematics education. This study also suggested that item writers and test developers should construct assessments where language is equally accessible for ELL students.

Keywords: English language learners, achievement gap, mathematics assessment, differential Item functioning, PISA

Edited by:
Dubravka Svetina Valdivia,
Indiana University Bloomington,
United States

Reviewed by:
Sebastian Weirich,
Institute for Educational Quality Improvement (IQB), Germany
Martin Hecht,
Humboldt University of Berlin, Germany

***Correspondence:**
Ruixue Liu
liuruixue2046@hotmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

***Received:** 22 January 2021*
***Accepted:** 30 April 2021*
***Published:** 27 July 2021*

Citation:
Liu R and Bradley KD (2021) Differential Item Functioning Among English Language Learners on a Large-Scale Mathematics Assessment.
Front. Psychol. 12:657335.
doi: 10.3389/fpsyg.2021.657335

INTRODUCTION

According to the U.S. Department of Education, ELLs are defined as students "who are being served in appropriate programs of language assistance" (National Center for Education Statistics, 2016). In recent decades, ELLs are a rapidly growing student group in the United States. The percentage of public-school students in the United States identified as ELLs grew from 3.8 million (8.1%) in 2000 to 4.8 million (9.5%) in 2015. In Fall 2015, the percentage of public-school students who were ELLs ranged from 1.0% in West Virginia to 21.0% in California (National Center for Education Statistics, 2018).

In school classrooms across the United States, ELLs are learning English and the content material in their required academic subjects simultaneously. To this end, ELLs have been found to lag behind their non-ELL peers on large-scale, standardized assessments, largely due to the high language demand in content areas, such as mathematics, science, reading comprehension, writing, and social studies (Abedi and Lord, 2001; Abedi et al., 2001; Abedi, 2002; Johnson and Monroe, 2004; Ockey, 2007; Mahoney, 2008; Walker et al., 2008; Martiniello, 2009; Wolf and Leon, 2009).

Purpose of Study

The current study aimed to explain the nature and potential sources of the gaps in mathematics achievement between ELLs and non-ELLs. The achievement gaps were examined using three differential item functioning (DIF) detection methods: Mantel-Haenszel (MH) procedure, Rasch model, and Hierarchical Generalized Linear Model (HGLM). Here, analyses were conducted at the item level instead of the traditional, total test result. Among the three methods, MH is the most widely used procedure to detect DIF in practice. The Rasch model (Rasch, 1960) allows for generalizability across samples and items, and it can identify poorly functioning items as well as unexpected responses. HGLM takes into account the nested structure of data where items are nested within students and students are nested within schools. At the student level, sources of DIF were investigated through the students' variations in mathematics self-efficacy, language proficiency, and socioeconomic status (SES). At the school level, school type and school educational resource were investigated as potential sources of DIF after controlling for the student variables.

The language of mathematics has been viewed as a unified system of meaning-making that incorporates multiple semiotics (Martiniello, 2009). The need to allocate cognitive resources to comprehend a problem presented in a non-primary language would reduce the resources available for the problem-solving process and result in increasing the probability of errors (Barbu, 2010). Therefore, it is necessary to investigate the assessment validity of mathematics for ELLs.

The Programme for International Student Assessment (PISA) is a large-scale assessment. The resulting data set allows researchers to investigate academic achievement and group membership from a variety of different viewpoints (Organisation for Economic Co-Operation Development, 2014). In this study, the U.S. sample of PISA 2012 was used. Seventy-six dichotomously coded items from PISA 2012 mathematics assessment were included to detect DIF effects.

Research Questions

Specifically, the present study will mainly address three research questions.

1) Do items from PISA 2012 mathematics assessment exhibit DIF between ELLs and non-ELLs for the U.S. sample?
2) If DIF is detected, can English language proficiency and other student characteristics (e.g., student SES, mathematics self-efficacy) explain DIF? That is, after controlling for these three student variables, whether DIF between ELLs and non-ELLs changes was examined.
3) If DIF is detected, can school type and school educational resources contribute to DIF?

The first research question aims to explore whether the PISA 2012 mathematics items are measuring in essentially the same way for ELLs and non-ELLs. The second and third research questions, incorporating a multilevel item analysis method, aim to identify the problem from multiple perspectives. Findings from this study can help educational researchers, administrators, and policymakers understand the nature of the achievement gap at item level instead of the total test level so the United States can be competitive in middle school mathematics education. This study also suggested that item writers and test developers should construct assessments where language is equally accessible for ELL students. The significance of this study lies in the empirical investigation of the gap between ELLs and non-ELLs in mathematics achievement at an item level and from the perspectives of both students and schools.

LITERATURE REVIEW

Assessment Validity for ELLs

Validity, as one of the most important attributes of an assessment, refers to how well the assessment tool measures the underlying outcome of interest. Validity is not a property of the tool itself, but rather of the interpretation or specific purpose of the assessment tool with particular settings and learners (American Educational Research Association et al., 2014). For ELLs, as well as for all populations, it is critical to consider whether the test scores can reflect the skill or proficiency that an assessment is intended to measure. Although students may have different English proficiency, the meaning of their scores on content assessments should be comparable (Educational Testing Service, 2009).

According to the Standard for Educational and Psychological Testing, *"the linguistic or reading demands of the test should be kept to the minimum necessary for the validity of assessment for the intended construct"* (American Educational Research Association et al., 2014, p. 82). Since ELL test takers have not yet acquired sufficient mastery of the English language, high language demand is still evident on mathematics assessment (Loughran, 2014). In this case, to a certain degree, the test unintentionally measures language proficiency. The lack of English proficiency used to comprehend mathematics assessment items results in an increased cognitive load and contributes to measurement error of ELL students' mathematics content knowledge (Educational Testing Service, 2009).

Differential Item Functioning

According to Holland and Wainer (1993), DIF analysis is a statistical technique to identify whether items on an assessment are of equal difficulty for examinees of different groups. DIF is present if an item on a test functions differently for different groups of interest (e.g., ELLs vs. non-ELLs), given the ability level. In the DIF analysis, examinees are matched based on their underlying ability (e.g., total score of an assessment), and differences in item performance between groups of examinees at the same level of ability are then determined.

DIF Detection Methods

There are numerous statistical methods for detecting DIF. The following methods are frequently used: Mantel-Haenszel procedure (Holland and Thayer, 1988), SIBTEST (Shealy and Stout, 1993), Item Response Theory methods (Camille and Shepard, 1994), logistic regression (Swaminathan and Rogers, 1990), and multilevel DIF analysis (Kamata, 2001). In this section, three methods applied in the current study were discussed in particular.

TABLE 1 | 2 × 2 × K contingency table.

	Scores on the Studied items		
	1 (Right)	**0 (Wrong)**	**Total**
Reference group	A_k	B_k	N_{rk}
Focal group	C_k	D_k	N_{fk}
Total	M_{1k}	M_{0k}	T_k

Mantel-Haenszel Procedure

The MH statistic was applied by Holland and Thayer (1988) in determining DIF. The MH DIF procedure compares dichotomous item performance between two groups after matching respondents on overall scores. Respondents in the focal and reference groups were matched on total test scores by dividing respondents in both groups into defined strata on those scores. The total scores were generated by summing item scores across all items. Estimates of the odds ratio for a given item can be calculated based on a 2 × 2 × K contingency table with k representing the *k*-th group ($k = 1,2,\ldots K$). **Table 1** shows the 2 × 2 contingency table for the *k*-th group of an item. The $A_{k,}$ B_k, C_k, and D_k denote the numbers of respondents in the cells. T_k represents the number of respondents in the k-the stratum.

The cells A_k and C_k represent the total number of respondents who answered the item correctly in the reference and focal groups, respectively, within the matched subgroup k. B_k and D_k denote the total number of respondents who answered the item incorrectly in the reference and focal groups, respectively, within subgroup k. N_{rk} and N_{fk} are the total number of respondents in the reference and focal groups, respectively, within k-th group. M_{1k} and M_{0k} denote the number of respondents who answered the item correct and incorrect, respectively, within k-th group.

The MH chi-square statistic is used for testing the null hypothesis of whether the population odds of getting an item correct is the same in the reference and focal groups. The statistic is given by

$$\text{MH CHISQ} = \frac{\left(\left|\sum_{k=0}^{n} \text{A}_k - \sum_{k=0}^{n} \text{E}\left(\text{A}_k\right)\right| - \frac{1}{2}\right)^2}{\sum_{k=0}^{n} \text{Var}(\text{A}_k)} \qquad (1)$$

In this equation, $\text{E}(\text{A}_\text{k})$ and $\text{Var}(\text{A}_\text{k})$ follow:

$$\text{E}(\text{A}_\text{k}) = \frac{N_{rk} * M_{1k}}{T_k} \qquad (2)$$

$$\text{Var}(\text{A}_\text{k}) = \frac{N_{rk} * N_{fk} * M_{1k} * M_{0k}}{\left[\,T_{K^2}\ \left(T_K - 1\right)\right]} \qquad (3)$$

The common odds-ratios formula is:

$$\text{OR}_{\text{MH}} = \sum_{k=0}^{n} \left[\frac{\text{A}_k * \text{D}_k}{\text{T}_k}\right] / \sum_{k=0}^{n} \left[\frac{B_k * C_k}{T_k}\right] \qquad (4)$$

The scale for OR_{MH} is from 0 to ∞, with $\text{OR}_{\text{MH}} = 1$ denoting the case of no DIF. For convenience, OR_{MH} is converted into a symmetrical scale $\Delta\ \text{OR}_{\text{MH}}$ given as

$$\Delta\text{OR}_{\text{MH}} = -2.35 ln(\text{OR}_{\text{MH}}) \qquad (5)$$

$\Delta\ \text{OR}_{\text{MH}}$ is applied as a measure of the DIF effect size. Educational Testing Service classified the DIF effect size as follows (Dorans and Holland, 1993) to aid in interpretation in applications: Class A denotes negligible magnitudes of DIF, when $|\Delta\ \text{OR}_{\text{MH}}| < 1.00$; Class B denotes moderate magnitudes of DIF, when $1.00 \leq |\ \Delta\ \text{OR}_{\text{MH}}| < 1.50$, and Class C denotes large magnitudes of DIF when $|\Delta\ \text{OR}_{\text{MH}}| \geq 1.50$.

The validity of the MH DIF detection method has been established by numerous studies (Qian, 2011). It is the most widely used procedure to detect DIF in practice since it is not only easy to understand and compute, it can provide both a significance test and estimate of the magnitude of DIF as well (Millsap, 2011). The major criticism of the MH procedure is the adequacy of using the total score as a substitute for the latent trait (Millsap, 2011). Besides, the requirement for sample sizes is a technical challenge to detecting items with DIF since DIF statistics become less stable as sample sizes decrease.

Rasch Model

Rasch model (Rasch, 1960) can produce a comprehensive and informative picture of the construct under measurement as well as the respondents on that measure. Since PISA 2012 employs the Rasch model to estimate student ability, item difficulty, and create the overall PISA literacy scale (Organisation for Economic Co-Operation Development, 2013), Rasch model was selected to detect DIF in the current study. Rasch model follows mathematically from the requirement of invariance of comparisons among persons and items (Andrich and Luo, 2003). Rasch model follows the following form:

$$P_{ij}(Y_{ij} = 1|\theta_j, b_i) = \frac{e^{(\theta_j\ -\ b_i\ )}}{\left[1 +\ e^{(\theta_j\ -\ b_i)}\right]} \qquad (6)$$

where p_{ij} is the probability of person j answering correctly to item i. θ_j is the person trait, or ability. b_i is the item parameter indicating difficulty of the item.

Rasch model provides a theoretically useful way to detect DIF which can be modeled using estimated item parameters and ability. A popular approach for detecting the DIF using the Rasch model is the Rasch separate calibration *t*-test method. This method is based on the differences between two separate calibrations of the same item from the subpopulation of interest, holding other item and person calibrations constant to ensure scale stability (Wright and Stone, 1979; Peabody and Wind, 2019). In Winsteps (Linacre, 2017), DIF detection using the Rasch model is conducted by a subtraction of the item location parameters (item difficulties) for two groups, d_1 and d_2. They are converted to standard normal variates using a pooled standard error. DIF is detected if the difference of item location parameters is statistically different.

$$t = \frac{d_1 - d_2}{\sqrt{var\ \left(d_1\right) + var\left(d_2\right)}} \qquad (7)$$

Hierarchical Generalized Linear Model

In behavioral and social sciences, data commonly have a nested structure. For example, repeated observations are nested within persons (e.g., responses nested within examinees), and persons are nested within organizational units such as classrooms, schools, and communities, and so on (Raudenbush and Bryk, 2002). Kamata (2001) proposed to use HGLM to detect DIF effects. Rasch model has been shown to be a special case of HGLM (Kamata, 2001; Raudenbush et al., 2003). In the three-level HGLM, the first level of the model is the item level, the second level the student level, and the third level is the school level. The three-level models have a nested structure where items are nested within students, and students are nested within schools.

Kamata (2001) mentioned that the three-level HGLM would be useful when the variation of the effect of a student characteristic variable across groups and the identification of a group-characteristic variable that explains such variation are of interest. There are several advantages of using HGLM to detect DIF in large-scale assessments. First, since the dependency of the data due to the nested data structure can be considered, DIF and item difficulty parameters can be modeled randomly across schools. Then the student and school variables can be examined simultaneously as potential sources of DIF. Second, additional student-level variables can be added as covariates to reduce student variations when identifying DIF. Third, various sources of DIF unique to each DIF item can be modeled simultaneously. Fourth, DIF detection using HGLM does not require two separate groups (focus and reference groups). This is especially beneficial if the source of the hypothesized DIF is a continuous variable (Qian, 2011).

METHOD

Data Source

The primary database used in this research is constructed from the PISA 2012. PISA is the most comprehensive and rigorous international assessment on 15-year-old students' performance in mathematics, reading, and science. According to the National Center for Education Statistics (2016), students in the PISA 2012 U.S. sample were born between July 1, 1996, and June 30, 1997. Finally, the PISA 2012 U.S. sample contains 4,978 students from 162 schools.

The targeted population for the focal group in this study is ELL students. Groups were identified using information collected from Student Questionnaire that was administered with the test. Basically, these two groups were designed to differ only in their relationships with English (as a first or second language). Home Language (ST25Q01) was used to form the groups. Home Language has the following binary categories: (1) language at home is the same as the language of the test and (2) language at home is another language. Students failing to answer this question were excluded from the current study. Finally, 670 students were identified as ELL students while 4,196 students were identified as non-ELL students.

MEASURES

Mathematics Items

In this study, 76 dichotomously coded items from PISA 2012 mathematics assessment were analyzed (See **Appendix A**). These items are either selected response multiple-choice or closed-constructed response. These items were scored as correct or incorrect and coded dichotomously with 1 and 0.

Student-Level Measures

The second research question aimed to investigate whether mathematics self-efficacy, English language proficiency, and student SES can explain DIF. These three variables were selected since they were found to be significant predictors to influence mathematics performance for ELLs (Aikens and Barbarin, 2008; Guglielmi, 2012). In addition, these variables served as control variables to reduce student variations. The measures of mathematics self-efficacy, language proficiency, and student SES were introduced in this section.

Mathematics Self-Efficacy

According to Bandura (1997), self-efficacy or perceived ability refers to the confidence an individual has in their ability to successfully perform a specific task. Previous studies indicated that mathematics self-efficacy and mathematics achievement were positively related. Students with high mathematics self-efficacy are associated with high mathematics achievement (e.g., Ayotola and Adedeji, 2009).

In PISA 2012, eight items were used to measure mathematics self-efficacy. These items ask students how confident they feel about having to do eight tasks (See **Appendix B**). Mathematics self-efficacy was measured in a four-point Likert-type scale (1 = Very confident; 2 = Confident; 3 = Not very confident; and 4 = Not at all confident). These items were scaled using IRT scaling methodology (Organisation for Economic Co-Operation Development, 2013).

Language Proficiency

When using HGLM to detect DIF between ELL and non-ELL students, language proficiency was used as one of the covariates. However, this information is not available in PISA 2012. Reading literacy, a proxy for language proficiency was used to represent language proficiency since understanding written text is the first form of language proficiency relevant to cognitive functions (Chen, 2010). PISA 2012 assessed reading literacy based on students' performance on three broad aspect categories including the ability to access and retrieve, integrate and interpret, and reflect and evaluate. These aspects were evaluated on printed and electronic texts which were defined as description, narration, exposition, argumentation, instruction, and transaction. In addition, IRT was used to estimate average scores for reading literacy (Organisation for Economic Co-Operation Development, 2014).

Socioeconomic Status

This study utilized the PISA index of economic, social, and cultural status (ESCS) to represent student SES. Variables comprising ESCS included home possessions (HOMEPOS), the number of books at home (HISEI), and the highest parental education expressed as years of schooling (PARED). The ESCS scores were obtained as component scores for the first principal component with zero being the score of an average OECD student and one being the standard deviation across equally weighted OECD countries. ESCS scores were calculated using the following formula:

$$\text{ESCS} = \frac{\beta 1^{*}\text{HOMEPOS} + \beta 2^{*}\text{HISEI} + \beta 3^{*}\text{PARED}}{\varepsilon} \tag{8}$$

where β_1, β_2, and β_3 are the OECD factor loadings and ε is the eigenvalue of the first principal component (Organisation for Economic Co-Operation Development, 2014).

School-Level Measures

The third research question aimed to investigate whether school type and school educational resources can contribute to DIF after controlling for student-level variables. These two variables were selected since they were found to be significant predictors to influence mathematics performance for ELLs (Freeman and Crawford, 2008; Han and Bridglall, 2009).

School Type

In PISA 2012, schools were categorized into public and private according to whether a private entity or a public agency has the ultimate power to make decisions concerning its affairs. The dummy variable of school type (SCHLTYPE) was created (0 = public, 1 = private).

School Educational Resources

The PISA 2012 school questionnaire contained 13 items about school educational resources, measuring principals' perceptions of potential factors hindering instruction at schools (e.g., a lack of qualified science teachers; shortage or inadequacy of science laboratory equipment; shortage or inadequacy of computer software for instruction; shortage or inadequacy of audio-visual resources). A four-point Likert-type scale was used (1 = Not at all, 2 = Very little, 3 = To some extent, 4 = A lot) (Organisation for Economic Co-Operation Development, 2014). The detailed items can be found in **Appendix C**. Responses to the 13 items measuring school educational resources were summed up and rescaled to a Z score to form the predictor of school educational resources.

STATISTICAL ANALYSES

Mantel-Haenszel Procedure

The PROC FREQ under the software of SAS 9.4 was used to conduct the MH procedure (Zhang, 2015). The total scores generated by summing item scores across all items were used to match students. Students' proficiency levels were controlled by stratifying students into five stratums based on the total scores. The DIF procedures in SAS 9.4 can provide key statistics including MH chi-square, common log-odds ratio, and estimated standard error. The MH chi-square statistic is distributed as chi-square with one degree of freedom. An alpha level of 0.01 was used for the MH procedure flag DIF items. The MH odds ratio is asymptotically normally distributed. ETS guidelines were used to classify items displaying DIF effects.

Rasch Model

The Rasch model was completed in the Winsteps measurement software, Version 3.9.1. Item difficulty measures (b_i in Equation 6) for both groups were calculated to examine whether the property of invariance was met. Winsteps outputs for DIF are equivalent to construct a "ruler" based on the persons, and measuring the items on it, first for the one person-group, then for the other person-group. In the output, the DIF contrast is the effect size of DIF and is a log-odds estimate. Specifically, the DIF contrast refers to the difference of item difficulty measures between ELLs and non-ELLs. A negative DIF contrast indicates that the item is more difficult for the ELLs.

Hierarchical Generalized Linear Model

The current study also used HGLM on DIF detection based on Kamata (2001) and Binici's (2007) studies. This study discussed that HGLM is equivalent to the Rasch model and showed how the two-level HGLM can be extended to a three-level latent regression model. Specifically, three models were created to answer the three research questions. Model 1 (DIF Identification Model) examined each item for DIF between ELLs and non-ELLs. Model 2 (DIF Estimation Model Controlling for Student-Level Variables) further examined whether student-level variables (mathematics self-efficacy, language proficiency, and SES) can explain DIF. Finally, Model 3 (Random Effects DIF Model) included student-level variables and school-level (school type and school educational resource) variables to explain DIF.

Model 1, including Level-1 and Level-2 models, was applied to the 76 items to detect DIF effects. Level-1 model is specified as given by Equation (9) where the log odds of the probability of answering each item correctly vs. incorrectly is a linear function of person ability and item difficulty.

$$Log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \sum_{q=1}^{76}\left(\beta_{qj}{*}X_{qij}\right) \tag{9}$$

X_{qij} is the q-th (q = 1, 2, …, 76) dummy coded variable that indicates the item i for student j. Its value is 1 when $q = i$ and 0 when $q \neq\ i$. β_{0j} is the effect of the reference item and β_{qj} is the difference between the q-th item and the reference item. The probability of student j getting an item i correct is noted as p_{ij}.

Level-2 model was then created by adding the group membership (ELL status) and modeling regression coefficients, βqj ($q = 1, 2,\ldots, 76$) in Equation (9) as given by Equation (10).

$$\begin{aligned}\beta_{0j} &= \tau_{00} + u_{0j}\\ \beta_{1j} &= \tau_{10} + \tau_{11}{*}ELL\\ &\ldots\\ \beta_{76j} &= \tau_{760} + \tau_{761}{*}ELL\end{aligned} \tag{10}$$

In Equation (10), coefficients τ_{11} to τ_{761} are the DIF coefficients associated with items 1 through 76. ELL students were coded as 1 and non-ELL students were coded as 0. A significant DIF coefficient indicates the existence of DIF for the item under investigation. The exponential term of the DIF coefficient is the odds of answering the corresponding item correctly by the reference vs. the focal group.

Model 2 further examined whether DIF effects decreases or disappears after controlling for student-level variables. As shown in Equation (11), Model 2 was created by adding student-level variables (language proficiency, SES, and mathematics self-efficacy) to Level-2 of Model 1. Level-1 of Model 2 is the same with Level-1 of Model 1 as shown as Equation (10). Level-2 of Model 2 was specified as follows.

$$
\begin{aligned}
\beta_{0j} &= \tau_{00} + u_{0j} \\
\beta_{1j} &= \tau_{10} + \tau_{11} * ELL \; + \; \tau_{12} * Language \; + \; \tau_{13} * SES + \tau_{14} * SE \\
\beta_{2j} &= \tau_{20} + \tau_{21} * ELL \; + \; \tau_{22} * Language \; + \; \tau_{23} * SES + \tau_{24} * SE \\
&\ldots \\
\beta_{76j} &= \tau_{760} + \tau_{761} * ELL \; + \; \tau_{762} * Language \; + \; \tau_{763} * SES \; + \tau_{764} * SE
\end{aligned} \tag{11}
$$

In Equation (11), coefficients τ_{11} to τ_{761} are the estimates of DIF after controlling for student-level variables. The exponential term of the DIF coefficients is the odds of answering the corresponding item correctly by the reference vs. the focus group. Coefficients τ_{12} to τ_{762} are the log odds of answering the corresponding item correctly with one unit of standard deviation (SD) increase in language proficiency. Similarly, τ_{13} to τ_{763} and τ_{14} to τ_{764} indicate the log odds of answering the corresponding item correctly with one unit of SD increase in SES and mathematics self-efficacy.

Model 3 examined whether school type and school educational resource contribute to the DIF. Specifically, Model 3 is a three-level DIF identification model including student-level and school-level variables. It investigated whether school type and school educational resources were significant predictors of DIF variations among 162 schools between ELL and non-ELL students.

DIF items that were detected by Model 1 were included in the analysis of Model 3. The Level-1, Level-2, and Level-3 of Model 3 were specified as shown in Equations (12, 13, and 14).

Level-1:

$$
\text{Log}\left(\frac{p_{ijk}}{1 - p_{ijk}}\right) = \beta_{0j} + \sum_{q=1}^{n} \left(\beta_{qjk} * X_{qijk}\right) \tag{12}
$$

Level-2:

$$
\begin{aligned}
\beta_{0jk} &= \tau_{00k} + u_{0jk} \\
\beta_{1jk} &= \tau_{10k} + \tau_{11k} * ELL \; + \; \tau_{12k} * Language \; + \; \tau_{13k} * SES + \; \tau_{14k} * SE \\
&\ldots \\
\beta_{njk} &= \tau_{n0k} + \tau_{n1k} * ELL \; + \; \tau_{n1k} * Language \; + \; \tau_{n1k} * SES \; + \; \tau_{n1k} * SE
\end{aligned} \tag{13}
$$

Level-3:

$$
\begin{aligned}
\tau_{00k} &= \pi_{000} + \varepsilon_{00k} \\
\tau_{10k} &= \pi_{100} \\
\tau_{11k} &= \pi_{110} + \pi_{111} * Schooltype + \; \pi_{112} * Resources \\
\tau_{12k} &= \pi_{120} \\
\tau_{13k} &= \pi_{130} \\
\tau_{14k} &= \pi_{140} \\
\tau_{20k} &= \pi_{200} \\
\tau_{21k} &= \pi_{210} + \pi_{211} * Schooltype + \; \pi_{212} * Resources \\
&\ldots \\
\tau_{100k} &= \pi_{1000} \\
\tau_{nk} &= \pi_{n10} + \pi_{n11} * Schooltype + \; \pi_{n12} * Resources \\
&\ldots
\end{aligned} \tag{14}
$$

The subscripts n and k indicate n-th DIF item and k-th school, respectively, at Level-3. At Level-2, coefficients of τ_{11k} to τ_{n1k} are random DIF coefficients that vary from school to school. At Level-3, coefficients π_{111} to π_{n11}indicate how much DIF increases when a school is from public to private, and coefficients π_{112} to π_{n12} denote how much DIF increases when school educational resources increase by one unit of SD. π_{120} to π_{n20} are the fixed regression coefficients for language proficiency. π_{130} to π_{n30} are the fixed regression coefficients for SES. π_{130} to π_{n30} are the fixed regression coefficients for mathematics self-efficacy.

The application of HGLM to detect DIF was conducted with PROC GLIMMIX under the software SAS 9.4. This procedure can fit models to outcome variables that generate a linear model with explanatory variables that account for variations at each level, utilizing variables specified at each level. PROC GLIMMIX can not only estimate model coefficients at each level, but it also predicts the random effects associated with each sampling unit at every level.

RESULTS

Mantel-Haenszel Procedure

The MH procedure was the first approach in this study to examine DIF effects between ELL and non-ELL students. **Table 2** shows the results of DIF effects using the MH procedure. Among the 76 items, 59 items with negligible DIF were categorized into Class A. Seven items with moderate values of $\Delta \text{OR}_{\text{MH}}$ were categorized into Class B. Ten items with large values of $\Delta \text{OR}_{\text{MH}}$ were categorized into Class C. In Class B, all seven items were in favor of non-ELL students. For example, Item 8 with the odds ratio of 0.55 indicated that ELL students are 45% less likely to answer this item correctly. In Class C, all ten items were in favor of non-ELL students. For example, Item 16 with the odds ratio of 0.50 indicated that ELL students are 50% less likely to answer this item correctly.

Rasch Model

DIF can be examined within the Rasch model by comparing item difficulties between groups. **Table 3** reports the difficulty estimates for both groups and their difficulty contrast. The t statistics were calculated using Equation (7). If the difficulty measures are significantly different between ELL and non-ELL students for the same item, this item was considered to have a DIF issue. A positive difficulty contrast indicates the item is more difficult for non-ELL students, and a negative difficulty contrast implies the item is more difficult for ELL students. According to **Table 2**, 14 items have been found to have DIF effects. Among the 14 items, nine items were more difficult for ELL students and five items were more difficult for non-ELL students. According to de Ayala (2009), items with difficulty contrast above 0.30 are considered as being noticeable. Thus these 14 items were found to display practically significant DIF effects.

Hierarchical Generalized Linear Model

Model 1

Table 4 summarizes the results of items with DIF effects using HGLM. Estimates are the DIF coefficients in Model 1. Estimates were exponentiated to obtain the DIF odds ratios. Then odds ratios were transformed to DIF effect size (Δ OR$_{\mathrm{MH}}$). Similar to the MH procedure, effect sizes were categorized into three classes. Among the 76 items, 66 items with negligible DIF were categorized into Class A. Five items with moderate values of ΔOR$_{\mathrm{MH}}$ were categorized into Class B. Six items with large values of ΔOR$_{\mathrm{MH}}$ were categorized into Class C. In Classes B and C, all the items were in favor of non-ELLs. For example, Item 8 with the DIF odds ratio of 0.62 indicated that ELL students are 38% less likely to answer this item correctly.

Model 2

In Model 2, student-level variables including mathematics self-efficacy, language proficiency, and SES were included to identify whether they are the potential sources of DIF between ELLs and non-ELLs. If the number of items showing DIF effects and their effect sizes decrease after controlling for the student-level

TABLE 2 | Summary results from the MH procedure to identify DIF effects.

Item number	MH chi-square	Odds-ratio	DIF effect size	Class
8. PM192Q01T	12.76**	0.55	1.42	B
16. PM420Q01T	21.23**	0.50	1.63	C
40. PM909Q01	14.84**	0.39	2.24	C
41. PM909Q02	20.02**	0.51	1.59	C
42. PM909Q03	31.67**	0.34	2.57	C
43. PM915Q01	13.10**	0.55	1.42	B
46. PM918Q02	23.31**	0.47	1.80	C
47. PM918Q05	7.78**	0.64	1.05	B
49. PM919Q02	10.10**	0.62	1.12	B
56. PM949Q01T	14.22**	0.56	1.35	B
57. PM949Q02T	14.00**	0.50	1.61	C
61. PM954Q02	16.74**	0.51	1.59	C
63. PM955Q01	16.90**	0.51	1.57	C
64. PM955Q02	11.75**	0.48	1.73	C
68. PM982Q04	9.52**	0.60	1.20	B
73. PM995Q02	5.11**	0.18	4.05	C

***$p \leq 0.01$.*

TABLE 4 | Summary of HGLM Model 1.

Item number	Estimates	Odds ratio	DIF effect size	Class
8. PM192Q01T	−0.47**	0.62	1.12	B
16. PM420Q01T	−0.52**	0.60	1.21	B
40. PM909Q01	−0.70**	0.50	1.65	C
42. PM909Q03	−0.87**	0.42	2.04	C
43. PM915Q01	−0.56**	0.57	1.32	B
46. PM918Q02	−0.64**	0.53	1.51	C
49. PM919Q02	−0.40**	0.65	1.01	B
61. PM954Q02	−0.64**	0.53	1.50	C
68. PM982Q04	−0.45**	0.64	1.05	B
73. PM995Q02	−1.60**	0.20	3.77	C

***$p \leq 0.01$.*

TABLE 3 | Summary of results from Rasch model to identify DIF effects.

Item number	Difficulty measures		Difficulty contrast	*t*	df
	Non-ELLs	ELLs			
8. PM192Q01T	0.25	0.66	−0.40*	−1.98	343
16. PM420Q01T	−0.88	−0.48	−0.40*	−2.28	394
27. PM564Q01	0.09	−0.48	0.48**	2.61	360
34. PM828Q02	−0.79	−1.37	0.58**	3.32	389
40. PM909Q01	−3.62	−3.08	−0.54*	−2.23	484
42. PM909Q03	0.82	1.56	−0.74**	−3.08	338
43. PM915Q01	−0.06	0.42	−0.48*	−2.45	326
46. PM918Q02	−2.04	−1.50	−0.53**	−2.98	464
49. PM919Q02	0.19	0.49	−0.30*	−1.67	416
51. PM923Q03	0.14	−0.41	0.54**	3.16	429
55. PM943Q02	4.69	3.66	1.03**	2.44	495
61. PM954Q02	0.65	1.21	−0.56*	−2.76	398
68. PM982Q04	0.01	0.36	−0.35*	−1.80	329
76. PM998Q04T	0.81	0.03	0.78**	4.19	385

**$p \leq 0.05$; **$p \leq 0.01$.*

TABLE 5 | Summary of results from HGLM Model 2.

Item number	Estimates	DIF Odds ratio	DIF effect size	Odds ratios of student-level variables		
				Mathematics self-efficacy	**Language proficiency**	**SES**
8. PM192Q01T	−0.46*	0.63	1.08	1.58**	1.01**	0.86
16. PM420Q01T	−0.36	0.70	0.85	1.06	1.01**	1.09
40. PM909Q01	−0.67*	0.51	1.58	1.02	1.01**	0.92
42. PM909Q03	−0.57*	0.56	1.34	1.03**	1.01**	1.21
43. PM915Q01	−0.46*	0.63	1.09	1.02**	1.01**	1.08*
46. PM918Q02	−0.59**	0.55	1.39	1.00	1.01**	0.99
49. PM919Q02	−0.42*	0.66	0.99	1.02	1.01**	0.91
61. PM954Q02	−0.50*	0.61	1.18	1.02**	1.01**	1.18*
68. PM982Q04	−0.28	0.75	0.66	1.18**	1.01**	1.14
73. PM995Q02	−1.46	0.23	3.43	1.18**	1.02**	1.11

*$p \leq 0.05$; **$p \leq 0.01$.

variables, these three variables are the potential sources of DIF at the student-level.

Table 5 displays the results from Model 2 to identify DIF effects controlling for student-level variables. Estimates are the DIF coefficients in Model 2. Odds ratios of student-level variables are the exponential terms for the regression coefficients of mathematics self-efficacy, language proficiency, and SES, which indicate the odds of getting each item correct associated one SD increase in those three variables.

Mathematics self-efficacy was a significant predictor on six of ten DIF items. For example, Item 8 with the odds ratio of 1.58 indicated that students with one SD increase of mathematics self-efficacy were 1.58 times more likely to answer this item correctly. Language proficiency was a significant predictor for all 10 items even its effect was minimal. For example, Item 73 with the odds ratio of 1.02 indicated that students with one SD increase of language proficiency were 1.02 times more likely to answer this item correctly. SES was a significant predictor on four of ten DIF items. For example, Item 61 with the odds ratio of 1.18 indicated that students with one SD increase of SES were 1.18 times more likely to answer this item correctly. After controlling for student-level variables, seven items still displayed DIF effects while the remaining three items no longer showed DIF effects. Besides, all the DIF effect sizes decreased after controlling for student-level variables.

Model 3

The three-level model was implemented for the 10 items displaying DIF effects. In this model, DIF effects were modeled to vary across 157 schools after controlling for the three student-level variables. Items with significant DIF variations across the schools were identified.

Table 6 displays the results from Model 3 to identify DIF effects controlling for student and school-level variables. Only odds ratios of school-level variables were displayed. Three out of 10 items were found to show significant DIF effects as both student and school-level variables were controlled (Item 8, 40, and 46). Nevertheless, school educational resources were not a significant predictor for these 10 items. School type was found to be a significant predictor for Item 40 and 46. For Item 40, students in private schools are 1.17 times more likely to answer this item correctly. For Item 46, students in private schools are 1.64 times more likely to answer this item correctly.

Consistency of Three DIF Detection Methods

Table 7 summarizes the DIF items identified by MH procedure, Rasch model, and HGLM. Seventeen items were identified with DIF effects in one of those three methods. Among those 17 items, eight items (Items 8, 16, 40, 43, 46, 49, 61, and 68) were identified with DIF effects by all three methods. Besides, the MH approach discovered six items with DIF effects that were not identified by the other two methods. Rasch model found five items with DIF effects that were not identified by the other two methods.

DISCUSSIONS

Summary of Findings

The first research question in this study asked whether 76 dichotomous items from PISA 2012 mathematics assessment exhibit DIF between ELLs and non-ELLs. In total, 21 items were identified with DIF effects by any of three methods. Sixteen items were found to be more difficult for ELLs while five items were more difficult for non-ELLs. Five items that were in favor of ELLs were identified by the Rasch model. Besides, eight items (Items 8, 16, 40, 43, 46, 49, 61, and 68) were identified with DIF effects by all three methods. These eight items were found to be more difficult for ELLs. the MH approach discovered six items with DIF effects which were not identified by the other two methods. The Rasch model found five items with DIF effects that were not identified by the other two methods.

Among the 10 DIF items that were identified by HGLM, seven items still displayed DIF effects after controlling for student-level variables. The rest of three items no longer showed DIF effects. These results suggest that mathematics self-efficacy, language proficiency, and SES are potential sources of DIF between ELLs and non-ELLs. Moreover, mathematics self-efficacy was a

TABLE 6 | Summary of results from HGLM Model 3.

Item number	Estimates	DIF odds ratio	DIF effect size	Odds ratios of school-level variables	
				School type	School educational resource
8. PM192Q01T	−0.66*	0.52	1.56	1.06	0.98
16. PM420Q01T	−0.45	0.64	1.06	0.70	0.94
40. PM909Q01	−0.60*	0.44	1.41	1.17*	0.98
42. PM909Q03	−0.56	0.57	1.31	0.76	1.07
43. PM915Q01	−0.46	0.63	1.09	1.08	0.97
46. PM918Q02	−0.78*	0.46	1.84	1.64*	0.94
49. PM919Q02	−0.17	0.84	0.40	0.97	0.97
61. PM954Q02	−0.25	0.78	0.59	0.82	1.00
68. PM982Q04	−0.51	0.60	1.19	1.06	0.95
73. PM995Q02	−1.37	0.25	3.22	1.01	0.96

*$p \leq 0.05$.

TABLE 7 | Summary of DIF items identified by three methods.

Item number	Mantel-Haenszel procedure	Rasch model	HGLM
8. PM192Q01T	Yes	Yes	Yes
16. PM420Q01T	Yes	Yes	Yes
27. PM564Q01		Yes	
34. PM828Q02		Yes	
40. PM909Q01	Yes	Yes	Yes
41. PM909Q02	Yes		
42. PM909Q03	Yes	Yes	Yes
43. PM915Q01	Yes	Yes	Yes
46. PM918Q02	Yes	Yes	Yes
47. PM918Q05	Yes		
49. PM919Q02	Yes	Yes	Yes
51. PM923Q03		Yes	
55. PM943Q02		Yes	
56. PM949Q01T	Yes		
57. PM949Q02T	Yes		
61. PM954Q02	Yes	Yes	Yes
63. PM955Q01	Yes		
64. PM955Q02	Yes		
68. PM982Q04	Yes	Yes	Yes
73. PM995Q02	Yes		Yes
76. PM998Q04T		Yes	

significant predictor on six of 10 DIF items. Language proficiency was a significant predictor for all 10 items even its effect was minimal. SES was a significant predictor on four of 10 DIF items. In addition, three items still displayed DIF effects after controlling for both student and school-level variables. The rest of seven items no longer displayed DIF effects. School type is a significant predictor for two items, while school educational resources were not a significant predictor for these 10 items.

Three DIF detection methods consistently identified eight items which were in favor of non-ELLs. This finding demonstrated the impact of English language proficiency on mathematics assessment, which aligned with some early studies. For instance, Abedi (2002) utilized existing data from several locations across the U.S. to examine the impact of students' language background on mathematics performance. The analyses mainly focused on the comparison between the level of performance of ELL and non-ELL students. The results discovered that ELLs generally perform lower than non-ELL students in mathematics. Similarly, Beal et al. (2010) found that the increase of mathematics test scores for the ELL students corresponded to English-reading proficiency in a non-linear manner. ELL students' English-reading proficiency predicted mathematics test scores, progress in the online mathematics tutorial, and mathematics self-concept.

Implications for Teachers and Educators

The current study can be used to inform mathematics teachers and educators on how best to respond to the instructional needs of their ELL students. In this study, mathematics self-efficacy was a significant predictor in six of 10 DIF items. Therefore, it is necessary for mathematics teachers and educators to develop this psychological belief for all students. Mathematics self-efficacy could be increased by using the right instructional strategies such as helping students to set learning goals, providing timely and explicit feedback, encouraging students to study harder, and using high achieving students as models (Liu and Hairy, 2009).

As a result of the language barrier and potentially negative perceptions of their academic ability from others, ELLs need additional support from mathematics teachers to enhance mathematics self-efficacy (Briscoe, 2014). Since the major sources of self-efficacy include mastery experience, vicarious experience, social persuasion, and psychological responses, it is helpful for ELLs to build self-efficacy by providing more successful experiences with mathematics, modeling, and verbal affirmations (Bandura, 1997).

The finding in the current study aligned with previous studies that proved language proficiency is a determinant factor to influence mathematics achievement (Abedi et al., 2001; Abedi, 2002; Haag et al., 2013; Loughran, 2014). It is suggested mathematics instruction should not isolate the word level from the discourse level. Teachers should provide opportunities for the discourse practices of explaining meanings of mathematical concepts and operations (Setati, 2005). The lexical support of meaning-related vocabulary offered in structured phrases rather than isolated words is important to the technical vocabulary (Prediger and Wessel, 2013).

Implications for Test Developers

The current study also provided some implications for test developers in terms of assessment development. Testing organizations should be more aware of linguistic diversity within the student population to make academic assessments more accessible for ELL students (Sireci and Faulkner-Bond, 2015). Moreover, quantitative and qualitative control procedures should be included to facilitate validity for subgroups of students. The quantitative process should include item analysis to evaluate statistical qualities such as item difficulty and discrimination. The qualitative process can incorporate the sensitivity review, which is an independent review of tests and items by experts trained to consider the unique characteristics of important subgroups (Sireci and Faulkner-Bond, 2015). ELL students can also be interviewed and asked to explain why the pilot items confused them (Ilich, 2013).

CONCLUSIONS

Examinations of DIF among language groups are a practical concern due to the increasing language diversity and the prevalence of testing. This study revealed that eight common items are identified with DIF effects using MH procedure, Rasch model, and HGLM. These eight items are all in favor of non-ELLs. These findings provided evidence supporting the claim that language ability has a negative impact on the mathematics performance for ELLs (Abedi, 2002; Martiniello, 2009; Loughran, 2014). Five items, identified by the Rasch model, was found to be in favor of ELLs. These items may be related to ELLs' prior educational experiences in their native languages. Although students are classified into ELLs as a result of their lack of English language proficiency, they may have been able to transfer key skills needed for those five items from their native languages to English.

When identifying the achievement gap between ELLs and non-ELLs, it is imperative to note that there are many possible reasons for the score differences. For instance, ELLs are more likely from low SES groups and may not have an equal chance to learn the content knowledge of mathematics. The unequal opportunities to learn result in true test score differences (Abedi et al., 2001). Inclusions of covariates in HGLM can solve this issue (Kamata, 2001). Finally, three items show strong evidence of DIF between ELLs and non-ELLs, even after controlling for student (e.g., mathematics self-efficacy, language proficiency, SES) and school (e.g., school type, school educational resources) level variables. Among the three items, two items (Item 40 and 46) with large DIF effect sizes (above 1.5) were categorized into Class C. According to ETS guidelines, items from Class C should not be used unless they are judged to be essential to meet test specifications (Zwick, 2012). Thus, it is suggested that PISA test developers should examine the language demand for these two items. Modifications or replacements should be made to reduce the DIF against ELLs.

The decreasing number of items showing DIF effects in HGLM Model 2 revealed that mathematics self-efficacy, language proficiency, and SES are potential sources of DIF between ELLs and non-ELLs. In addition, the number of DIF items continued to decrease after controlling for both student and school-level variables. This finding implied that DIF effects between ELLs and non-ELLs can vary in different schools. School type and school educational resources were also potential sources of DIF effects.

Since it is difficult to estimate the amount of error in the data (e.g., missing data) from empirical studies, applying more than one DIF detection approach was suggested to increase the confidence in the results (Hidalgo and LÓPez-Pina, 2004). There is some disagreement among three DIF detection approaches. This disagreement mainly resulted from the different mechanism of DIF detection methods. For instance, MH procedure used raw scores to match students from different groups for DIF detection, but raw scores cannot represent students' true ability levels properly when tests have DIF items or the impact is large (Jin et al., 2018). Besides, the MH procedure failed to make any assumptions about the classical test theory decomposition of scores. By comparison, Rasch model and HGLM can be classified into the parametric and the latent matching category. They are closely linked to a test theory that decomposes an observed score into a systematic true score and a stochastic error score (Kim, 2003). In terms of the Rasch model and HGLM, although the Rasch model has been regarded as a special case of HGLM (Kamata, 2001; Raudenbush et al., 2003), their detection mechanism of DIF effects is different. A general approach for detecting the DIF using the Rasch model is the Rasch separate calibration *t*-test method while HGLM relies on evaluating the interaction effect of item by person characteristic variable. Additionally, Kamata (2001) suggested that HGLM can be carried out by using one of the items as the reference item. As a result, within HGLM analyses, results may slightly vary due to the selection of the reference item.

For practitioners, it is important to note the roles of statistical DIF and substantive DIF (Peabody and Wind, 2019). All of the existing DIF detection methods are designed to identify statistical DIF. The analyst is expected to identify as many as possible to ensure that all items exhibit real DIF. After all items with the statistical DIF are identified, content experts will determine if substantive DIF exists.

LIMITATIONS AND FUTURE RESEARCH

In this study, students whose primary language spoken at home was not English were classified as ELL students. While reading literacy was used to represent language proficiency and was controlled as a covariate in HGLM, levels of English proficiency for ELL students were unknown. Some students who were categorized into the ELL group may have transitioned out of English as a second language class. This variability within the ELL group may limit the results. It is recommended that large-scale assessments (e.g., PISA) can collect samples of data from students who are at varying levels of English language proficiency.

Potential sources of DIF between ELLs and non-ELLs should be interpreted with cautions. This study relied on correlational analysis to ascertain the relationship between DIF effects and other covariates. Further casual explanation for the achievement gap between ELLs and non-ELLs needs to be investigated at both student and school levels.

PISA only released a small portion of mathematics items so that it is impossible to review the detailed content of each item. Further comprehensive content analysis on the DIF items should be conducted when the PISA 2012 mathematics assessment items are released in the future. First, vocabulary and terminologies of DIF items should be reviewed by mathematics educators and assessment experts to see whether cultural bias exists. Second, content reviews can be made to rate the level of linguistic complexity by experts in the areas of literacy, linguistics, and bilingual education. Whether linguistic complexity can predict the magnitude of DIF effects between ELL and non-ELL students can be investigated.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm.

ETHICS STATEMENT

The data included in this study was derived from PISA. Therefore, the study did not require ethics approval.

AUTHOR CONTRIBUTIONS

RL: generation of research goals and aims, development of research methodology, data analysis, and writing the draft. KB: providing mentorship on the research, reviewing, and editing the draft. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.657335/full#supplementary-material

REFERENCES

Abedi, J. (2002). Standardized achievement tests and English Language learners: psychometrics issues. *Educ. Assess.* 8, 231–257. doi: 10.1207/S15326977EA0803_02

Abedi, J., Hofstetter, C., Baker, E., and Lord, C. (2001). *NAEP Math Performance and Test Accommodations: Interactions with Student Language Background (CSE Tech. Rep. No. 536).* Los Angeles, LA: National Center for Research on Evaluations, Standards, and Student Testing.

Abedi, J., and Lord, C. (2001). The language factor in mathematics tests. *Appl. Measure. Educ.* 14, 219–234. doi: 10.1207/S15324818AME1403_2

Aikens, N. L., and Barbarin, O. (2008). Socioeconomic differences in reading trajectories: the contribution of family, neighborhood, and school contexts. *J. Educ. Psychol.* 100, 235–251. doi: 10.1037/0022-0663.100.2.235

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing.* Washington, DC: AERA.

Andrich, D., and Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *J. Appl. Meas.* 4, 205–221.

Ayotola, A., and Adedeji, T. (2009). The relationship between mathematics self-efficacy and achievement in Mathematics. *Proc. Soc. Behav. Sci.* 1, 953–957. doi: 10.1016/j.sbspro.2009.01.169

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control.* New York, NY: Freeman.

Barbu, O. (2010). Effects of linguistic complexity and math difficulty on word problem solving ability by English students. *Int. J. Educ.* 2, 1–19. doi: 10.5296/ije.v2i2.508

Beal, C. R., Adams, N., and Cohen, P. R. (2010). Reading proficiency and mathematics problem solving by English Learners. *Urban Educ.* 45, 58–74. doi: 10.1177/0042085909352143

Binici, S. (2007). Random effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: A comparison of estimation methods. (Unpublished doctoral dissertation). Florida State University.

Briscoe, F. M. (2014). "The biggest problem": school leaders' covert construction of Latino ELL families: Institutional racism in a neoliberal schooling context. *J. Lang. Ident. Educ.* 13, 354–373. doi: 10.1080/15348458.2014.958041

Camille, G., and Shepard, L. (1994). *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage.

Chen, F. (2010). *Differential Language Influence on Math Achievement.* Greensboro, NC: ProQuest Dissertations and Theses database.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory.* New York, NY: Guilford.

Dorans, N. J., and Holland, P. W. (1993). "DIF detection and description: Mantel-Haenszel and standardization," in *Differential Item Functioning*, eds P. W. Holland and H. Wainer (Hillsdale NJ: Erlbaum).

Educational Testing Service (2009). *Guidelines for the Assessment of English Language Learners.* Available online at: https://www.ets.org/s/about/pdf/ell_guidelines.pdf

Freeman, B., and Crawford, L. (2008). Creating a middle school mathematics curriculum for English-language learners. *Remed. Spec. Educ.* 29, 9–19. doi: 10.1177/0741932507309717

Guglielmi, R. S. (2012). Math and science achievement in English language learners: multivariate latent growth modeling of predictors, mediators, and moderators. *J. Educ. Psychol.* 104, 580–602. doi: 10.1037/a0027378

Haag, N., Hept, B., Stanat, P., Kuhl, P., and Pant, H. A. (2013). Second language learners' performance in mathematics: disentangling the effects of academic language features. *Learn. Instruct.* 28, 24–34. doi: 10.1016/j.learninstruc.2013.04.001

Han, W., and Bridglall, B. L. (2009). Assessing school supports for ELL students using the ECLSK. *Early Child. Res. Q.* 24, 445–462. doi: 10.1016/j.ecresq.2009.08.003

Hidalgo, M. D., and LÓPez-Pina, J. A. (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educ. Psychol. Meas.* 64, 903–915. doi: 10.1177/0013164403261769

Holland, P., and Thayer, T. (1988). "Differential item functioning and the Mantel-Haenszel procedure," in *Test Validity,* eds H. Wainer and H. I. Braun (Hillsdale, NJ: LEA).

Holland, P., and Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ilich, M. (2013). Differential Item Functioning (DIF) among Spanish-Speaking English Language Learners (ELLs) in State Science Tests. (Unpublished doctoral dissertation). University of Washington, Seattle, WA, USA.

Jin, K., Chen, H., and Wang, W. (2018). Using odds ratios to detect differential item functioning. *Appl. Psychol. Measure.* 42, 613–629. doi: 10.1177/0146621618762738

Johnson, E., and Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assess. Effect. Interven.* 29, 35–45. doi: 10.1177/073724770402900303

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *J. Educ. Measure.* 38, 79–93. doi: 10.1111/j.1745-3984.2001.tb01117.x

Kim, W. (2003). Development of a differential item functioning (DIF) procedure using the hierarchical generalized liner model: a comparison study with logistic regression. (Unpublished doctoral dissertation). Pennsylvania State University, State College, PA, United States.

Linacre, J. M. (2017). *A User's Guide to Winsteps, Ministep, Rasch-Model Computer Programs: Program Manual 4.0.0*. Available online at: http://www.winsteps.com/a/winsteps-manual.pdf

Liu, X., and Hairy, K. (2009). The effect of mathematics self-efficacy on mathematics achievement of high school students. Paper presented at the annual conference of Northeastern Educational Research Association. Rocky Hill, Connecticut, USA.

Loughran, J. (2014). *Understanding differential item functioning for English Language Learners: The influence of linguistic complexity features (Unpublished doctoral dissertation)*. University of Kansas, Lawrence, KS, USA.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the National Assessment of Educational Progress. *Int. J. Test.* 8, 14–33. doi: 10.1080/15305050701808615

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English Language Learners in math tests. *Educ. Assess.* 14, 161–170. doi: 10.1080/10627190903422906

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge. doi: 10.4324/9780203821961

National Center for Education Statistics (2016). *Methodology and Technical Notes-Sampling and Data Collection in the United State*. Available online at: https://nces.ed.gov/surveys/pisa/pisa2012/pisa2012highlights_9b.asp

National Center for Education Statistics (2018). *English Language Learners in Public Schools*. Available online at: https://nces.ed.gov/fastfacts/display.asp?id=96

Ockey, G. J. (2007). Investigating the validity of math word problems for English language earners with DIF. *Lang. Assess. Q.* 4, 149–164. doi: 10.1080/15434300701375717

Organisation for Economic Co-Operation and Development (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing.

Organisation for Economic Co-Operation and Development (2014). *PISA 20012 Technical Report*. Paris: OECD Publishing.

Peabody, M., and Wind, S. (2019). Exploring the stability of differential item functioning across administrations and critical values using the Rasch separate calibration t-test method. *Measure. Interdisc. Res. Perspect.* 17, 78–92. doi: 10.1080/15366367.2018.1533782

Prediger, S., and Wessel, L. (2013). Fostering German language learners' constructions of meanings for fractions – Design and effects of a language- and mathematics integrated intervention. *Mathe. Educ. Res. J.* 25, 435–456. doi: 10.1007/s13394-013-0079-2

Qian, X. (2011). *A Multi-Level Differential Item Functioning Analysis of the Trends in International Mathematics and Science Study (TIMSS) 2003: Potential Sources of Race/Ethnicity and Gender Differences Among U.S. 8th Graders' Science Achievement*. Newark, DE: ProQuest Dissertations and Theses database.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: The University of Chicago Press.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd ed.)*. Newbury Park, CA: Sage.

Raudenbush, S. W., Johnson, C., and Sampson, R. J. (2003). A multivariate, multilevel Rasch model with application to self–reported criminal behavior. *Sociol. Methodol.* 33, 169–211. doi: 10.1111/j.0081-1750.2003.t01-1-00130.x

Setati, M. (2005). Teaching mathematics in a primary multilingual classroom. *J. Res. Math. Educ.* 36, 447–466.

Shealy, R., and Stout, W. (1993). A model-based standardization approach that separates true Bias/DIF from group ability differences and detects test Bias/DTF as well as item Bias/DIF. *Psychometrika* 58, 159–194. doi: 10.1007/BF02294572

Sireci, S. G., and Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Rev. Res. Educ.* 39, 215–252. doi: 10.3102/0091732X14557003

Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *J. Educ. Measure.* 27, 361–370. doi: 10.1111/j.1745-3984.1990.tb00754.x

Walker, C. M., Zhang, B., and Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Appl. Measure. Educ.* 21, 162–181. doi: 10.1080/08957340801926201

Wolf, M. K., and Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educ. Assess.* 14. 139–159. doi: 10.1080/10627190903425883

Wright, B. D., and Stone, M. H. (1979). *Best Test Design*, Chicago, IL: MESA Press.

Zhang, Y. (2015). *Multiple Ways to Detect Differential Item Functioning in SAS (2900-2015)*. Educational Testing Service. Available online at: https://support.sas.com/resources/papers/proceedings15/2900-2015.pdf

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*. Available online at: https://www.ets.org/Media/Research/pdf/RR-12-08.pdf

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

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