Running the Wrong Race? The Case of PISA for Development

DAVID RUTKOWSKI AND LESLIE RUTKOWSKI

With the aim of developing an assessment more appropriate for low-income countries, in 2013 the Organisation for Economic Co-operation and Development initiated the Programme for International Student Assessment for Development (PISA-D). Designed as a one-off assessment project resulting in scores comparable to PISA, nine countries joined the project. In this article, we focus on whether PISA-D does what it sets out to do, namely, appropriately measure the specified content in the selected populations. In particular, we investigate the degree to which PISA-D is well-suited for measuring low-performing educational systems. To that end, we detail how well the modifications designed to make PISA-D an "easy" instrument capture assessed populations. We conclude the article with policy implications of our findings.

Introduction

The Organisation for Economic Co-operation and Development (OECD) has emerged as an influential organization in global education largely though its collection of educational indicators. The OECD developed and maintains an elaborate evaluation system collecting educational data on both member and nonmember systems from early childhood to 65-year-olds. The most extensive and well-known of the OECD educational evaluations is the Programme for International Student Assessment (PISA), originally administered in 2000 to assess the 28 (at that time) OECD member countries' 15-year-olds' ability in reading, mathematics, and science (OECD 2000). Although PISA participation grew from 43 educational systems in 2000 to 79 in 2018, only OECD members and Brazil (a PISA associate) are able to vote on the PISA governing board, which determine PISA's policy priorities and oversees major decisions made about the assessment (OECD, n.d.). Given PISA's history and governance structure, it should be no surprise that PISA is largely geared toward the evaluation of its economically developed members. Regardless of the historic focus on OECD countries, currently PISA participants from non-OECD systems outnumber OECD member countries. As demonstrated in Rutkowski

Received November 22, 2019; revised June 15, 2020; accepted September 16, 2020; electronically published March 5, 2021

Comparative Education Review, vol. 65, no. 1.

0010-4086/2021/6501-0008$10.00

et al. (2019), cross-cultural differences in PISA 2015 achievement are large, suggesting that PISA may not be ideally suited to measure all participating countries equally well. Naturally, this poses challenges for the organization, namely, how to create an assessment for a heterogeneous set of countries when the structure is designed for and governed by a select group of relatively homogeneous countries.

The OECD, aware for some time of the widening proficiency differences between participants, has attempted to make various accommodations for lower-performing countries. For example, in 2009 the OECD included sets of easy booklets into its design to help better assess lower achieving countries (OECD 2010). Enthusiasm for this innovation was substantial, with 20 educational systems selecting this option (OECD 2012). However, as shown in Rutkowski et al. (2018), the approach of including easy booklets did little to improve measurement when compared to administering a common test for all participants. In other words, although the OECD attempted to accommodate low-performing countries the modification did not improve measurement for low performers.

With the aim of developing an assessment more appropriate for lower preforming countries, in 2013 the OECD initiated PISA for Development (PISA-D) with a stated purpose "to make the assessment more accessible and relevant to low-to-middle-income countries" (OECD 2016a, 2). Originally designed as a one-off project resulting in scores comparable to PISA, nine countries are members of the PISA-D project including: Bhutan, Cambodia, Ecuador, Guatemala, Honduras, Panama, Paraguay, Senegal, and Zambia. PISA-D includes three technical strands: Strand A, enhancement of PISA's cognitive instruments; Strand B, enhancement of PISA's contextual questionnaires; and Strand C, the development of an approach and methodology for incorporating out-of-school 15-year-olds in the assessment (Carr-Hill 2015). In the current article we limit our emphasis to Strand A. This is partly due to the fact that Strand C data and technical documentation was not available as of the writing of this article. Furthermore, our research question does not involve the data, frameworks, or technical documentation for Strand B. As such, we focus on the seven countries that assessed the in-school population (Bhutan and Panama only took part in the out-of-school assessment) to investigate whether the assessment is suited for measuring a group of lower- and middle-income countries.

Data collection for the in-school assessments were completed in 2017 and results were reported in 2018. According to the OECD (2016a, 3) the assessment had four aims:

- Provide policy makers in the participating countries with insights on how to help students learn better, teachers to teach better, and school systems to operate more effectively.

- Help to build the capacity of participating countries to conduct large-scale learning assessments, and analyze and use the results to support national policies and evidence-based decision making.
- Enhance PISA to make it more relevant to a wider range of countries and thus enable greater PISA participation by middle- and low-income countries.
- Contribute to the monitoring and achievements of the Education Sustainable Development Goal, which emphasizes quality and equity of learning outcomes for children, young people and adults.

The assessment is marketed as a gateway to PISA participation where countries are provided with resources to help them gain the skills needed to fully engage in large-scale educational assessments (OECD 2016b). Furthermore, participation in international large-scale assessments (ILSAs) such as PISA-D has the benefit of building national assessment capacity (Lockheed et al. 2015). As such, PISA-D expands the reach of PISA and demonstrates a push by the OECD to create an assessment system and way of envisioning educational systems that reaches beyond only rich OECD members, to an extended group of countries.

Similar to much of the critical research around PISA (Meyer and Benavot 2013; Sjøberg 2015; Engel et al. 2019), a small but growing literature around PISA-D has emerged. For example, Addey and Sellar (2018) have voiced concerns, noting that PISA-D affords the OECD greater educational governance in national systems. Auld et al. (2019) argue that Cambodia was a less than willing participant in the assessment and that the OECD, World Bank, and UNESCO's drive to assess educational quality eclipsed national needs for international benchmarking. These authors suggest that PISA-D is a social experiment, affording the OECD a leadership role in evaluating (e.g., judging the merit and worth) of low-performing educational systems. In this article, however, we take a step back from commentary on the role of PISA-D in society and focus on whether PISA-D does what it sets out to do, namely, appropriately measure the specified content of the selected populations. Our line of inquiry aligns with concerns posed by Adams and Cresswell (2016) in an OECD report that cautions against an overreliance by PISA-D countries on the extant PISA framework and questions. With this in mind, we investigate the degree to which PISA-D is well suited for measuring low-performing educational systems. To that end, we detail how well the modifications designed to make PISA-D an "easy" instrument capture assessed populations. We borrow the concepts of construct mapping from Wilson (2004) and the notion of proficiency as defined by PISA (Kirsch et al. 2002; OECD 2017a). As a second emphasis of this article, we discuss the importance of equivalent cross-cultural measurement and problems, especially in low-performing countries, with the method used to evaluate measurement equivalence in PISA-D. This has an important consequence that it is a further challenge to detect whether the assessment is functioning as intended across all measured populations.

PISA-D In-School Population

According to the OECD, the PISA-D content framework, which determines what would be assessed, was overseen by the PISA Governing Board (note: participating PISA-D countries cannot vote on this board). The OECD claims that the PISA-D framework is an extension of the PISA framework, which attempts to assess whether students "can extrapolate from what they have learned and apply their knowledge in new situations" on three content domains: reading, mathematics, and science (OECD 2019c, chap. 1, p. 2). From the seven participating countries that assessed their in-school populations, a total sample of 34,605 students were given the paper-based assessment. In addition to the cognitive portion of the assessment, context questionnaires were administered that focus on student's home, family, and school background; school organization and education provisions in school; and teaching practices. Each student was allotted 2 hours to complete the assessment and an additional 35 minutes for the background questionnaire (OECD 2019c).

According to the OECD (2019c, chap. 2, p. 1), the cognitive assessment included:

- A compulsory assessment of reading, mathematics, and science, with equal weights for each of the three domains (i.e., no major/minor domain distinction as is made in PISA).
- Paper-based cognitive instruments linked to PISA. This meant that a majority of items were selected from previous cycles of PISA but complemented with existing materials from surveys including PISA for Schools, the Programme for the International Assessment of Adult Competencies (PIAAC), the World Bank's Skills Toward Employability and Productivity (STEP) assessment, and the Literacy Assessment and Monitoring Program (LAMP).
- No new cognitive items.
- Items that were reviewed and selected to meet the measurement goals of PISA-D.

Notable from this list is the absence of new items and that the OECD choose to take most items from previous rounds of PISA—an assessment deemed too difficult for these populations. Specifically, the OECD (2019c) explains that PISA 2015 trend items were the primary source of PISA-D items ranging across the proficiency continuum. Using trend items enabled PISA-D results to be placed on the main PISA scale. To ensure that the test made some attempt to focus on lower performers, approximately 60 percent of all items selected represented the PISA proficiency level 2 or below (6 being the highest).

Similar to PISA, PISA-D incorporated a rotated booklet design with multiple matrix sampling where every student was only administered a portion of the test. This design necessitates that special methods are used to estimate achievement, commonly referred to as *plausible value methods* (von Davier et al. 2009). Essentially, these methods produce multiple achievement values for each student that are a random draw from a model-based distribution of achievement. Although this reduces individual testing burden, a limitation, as in main PISA, is that because of sampling and the assessment design, school and student level interpretation of the results are prohibited. The main survey was broken down into 16 30-minute clusters (four for each domain) where 50–60 percent of each cluster included trend items from PISA 2015 and 40–50 percent were items from other assessments. As can be seen in table 1, the 16 clusters were separated into 12 booklets that included clusters representing two domains.

For PISA-D Strand A (the focus of this article), the target population was 15-year-olds attending educational institutions in grade 7 or higher, excluding those schooled at home, in the workplace, or out of the country. As shown in table 2, in both Cambodia and Honduras, less than half of the 15-year-olds were considered enrolled in an educational institution.

The study employed a two-stage stratified sampling design, with the first-stage including individual schools with PISA-D eligible students and the second stage including eligible students within the sampled school. Full sampling details are available in the technical documentation (OECD 2019c). Similar to PISA, PISA-D reports a coverage index, which is the proportion of the national population of 15-year-olds covered by the student sample. The proportion is "obtained by dividing the number of students represented by the PISA-D sample . . . by the total number of 15-year-olds estimated from demographic projections" (MoEYS 2018, 17). As shown in table 2, Cambodia's and Senegal's coverage index were 28 percent and 29 percent, respectively,

TABLE 1
PISA-D Main Assessment Design

	Cluster					
Booklet	RC	1	2	RC	3	4
1	RC1	R1	R2		S1	S2
2		S2	S3	RC2	R2	R3
3	RC3	R3	R4		S3	S4
4		S4	S1	RC4	R4	R1
5		S1	S2		M1	M2
6		M2	M3		S2	S3
7		S3	S4		M3	M4
8		M4	M1		S4	S1
9		M1	M2	RC1	R1	R2
10	RC2	R2	R3		M2	M3
11		M3	M4	RC3	R3	R4
12	RC4	R4	R1		M4	M1

Source.—OECD 2019b, chap. 2, p. 5.

Note.—R1–R4 are reading literacy clusters; RC1–RC4 are reading components clusters; M1–M4 are mathematical literacy clusters; S1–S4 are scientific literacy clusters.

TABLE 2
PISA-D Countries Sampling Frame

	All 15-Year-Olds	Enrolled 15-Year-Olds	Coverage Index	GDP per Capita (USD)
Cambodia	370,856	166,144	0.28	1,510
Ecuador	352,702	300,364	0.61	6,345
Guatemala	387,167	199,582	0.47	4,549
Honduras	193,268	93,767	0.41	2,500
Paraguay[a]	135,869	100,542	0.56	5,821
Senegal	337,636	257,384	0.29	1,522
Zambia	361,058	193,637	0.36	1,540

Source.—OECD 2019b.
Note.—GDP = gross domestic product.
[a]Coverage index may be significantly underestimated in Paraguay.

compared to a PISA-D average of 43 percent and a PISA 2015 OECD average of 89 percent (OECD 2019c, 207). In other words, although the OECD claims to have met its sampling goals, PISA-D represents less than half of 15-year-olds in all but one participating county. In light of these differences in population representation, caution is urged when drawing inferences from PISA-D results, particularly when comparing these countries to main PISA participants. We now turn to how well PISA-D measured the sampled students.

A Look at Test-Examinee Match in PISA-D

Following a similar approach to Rutkowski et al. (2019) we examine the degree to which PISA-D is well matched to participating populations using a visual means to relate examinees to items, referred to as a construct map (Wilson 2004). A construct map shows the distribution of examinee proficiency against the item location on the same continuum. To develop these maps, we rely on a principle in item response theory (IRT) that allows us to place test items and examinees on the same scale (Embretson and Reise 2000). This offers the possibility of comparing individuals to one another, items with one another, and comparing individuals with items. For our construct maps, we use the PISA scale (historical achievement mean of 500 and standard deviation of 100). An item's location—in educational measurement terms—is the point along the achievement continuum where an examinee that is at the same location has some probability of a correct answer. In line with the OECD's approach, we use a correct response probability of .62, or RP62. Items that are higher on the PISA scale are more difficult than items that are lower on the scale. For example, an item that is located at the OECD average (500) is more difficult than an item that is located at 480 and less difficult than an item that is located at 520. In a similar vein, examinees that are located at 500 are said to be more proficient than examinees at 480 and less proficient than examinees at 520 on the scale. To plot the proficiency distribution for each country, we use the first plausible value. Such a representation gives a clear picture of the degree to which a group of examinees are matched to a test. These graphical representations give an overall picture of the alignment between a group of examinees and the test. A test that is well matched to the examinees is one where the items are located at or around substantial portions of the proficiency distribution. Gaps in item locations indicate that the construct is not well measured for those areas of the proficiency continuum.

We can also think about these results in terms of the probability of a correct response for respondents to items of average difficulty. To do so, we select a typical math item (with a difficulty at about 500, the historic mean). The item that most matches this description is a multiple choice item,[1] taken from PISA for Schools. This item deals with the content domain uncertainty and data and requires students to formulate a response. Then, we select hypothetical students from various points along the proficiency continuum, to include a typical student (at the country's average), a high achieving student (one standard deviation above the country mean), and a low achieving student (one standard deviation below the country mean). Although these values are somewhat arbitrary, they provide insights into the degree to which students from various points along the proficiency continuum can reasonably be expected to engage with the test. We use the standard IRT equation that relates a person and an item to the probability of a correct response, given as:

$$P(x_j = 1|\theta_i, a_j, b_j) = \frac{1}{1 + e^{a_j(\theta_i - b_j)}}, \tag{1}$$

where θ_i is the proficiency level for examinee *i*. The parameters a_j and b_j are characteristics of item *j*. In particular, a_j indicates the degree to which item *j* can discriminate among students with different proficiency. Higher values indicate that the item discriminates well. And b_j is the item difficulty, which locates item *j* along the proficiency continuum and can be interpreted as the proficiency value that corresponds to a 50 percent chance of a correct answer. Higher values indicate a more difficult item, lower values indicate an easier item. Our selected item has $a_j = 1.002$ and $b_j = -.378$. A comprehensive description of these parameters and their interpretations are well outside the scope of this article; however, interested readers are encouraged to consult Hambleton and Swaminathan (1985) or Embretson and Reise (2000) for accessible introductions to IRT. Suffice to say, however, that this item is relatively easy and moderately discriminating.

We start by presenting an overall picture of proficiency distributions for all seven PISA-D countries set against the item locations for each content domain, located in figure 1*A*–*C*. In these plots, we can see that, for all but the highest-performing PISA-D countries, there is little overlap between the country's achievement distribution and the item locations. Next, we present

[1] In the documentation, this item can be located by its ID, PM5104Q01.

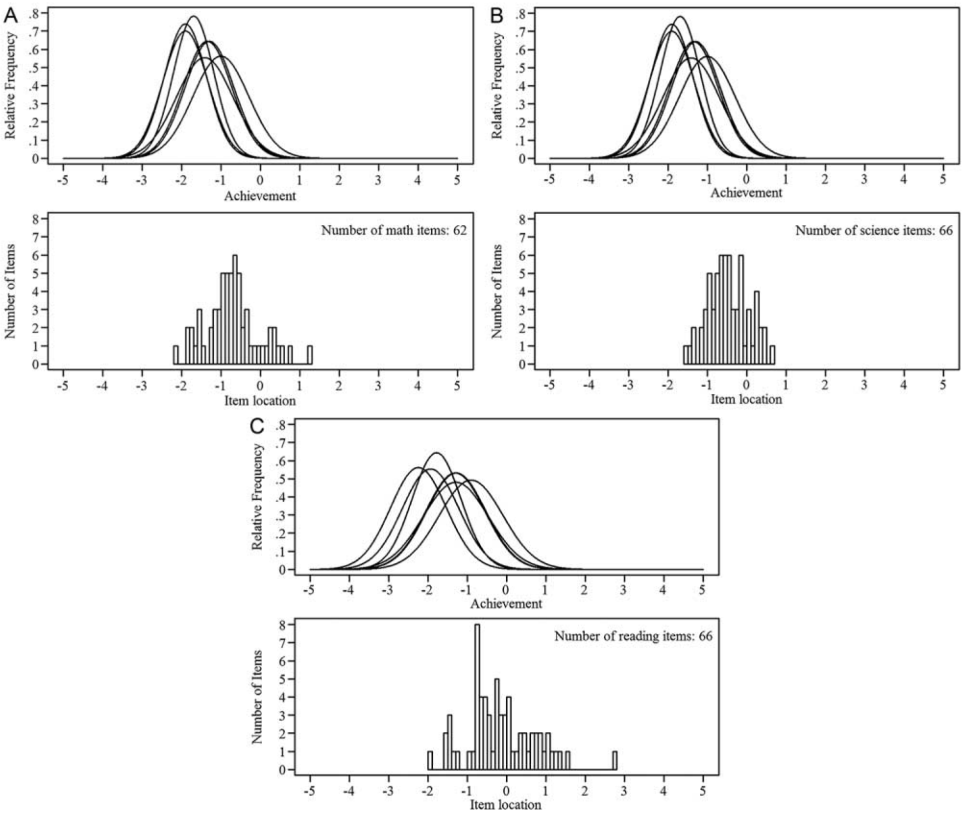

Fig. 1.—Relative frequency and item location for PISA-D participating countries

the construct maps, country by country, for mathematics (fig. 2*A*–*G*). For the country-by-country construct maps, the *x*-axis in each plot represents the PISA scale. From the top to the bottom panels within each figure, the *y*-axes represent (*a*) a sample-weighted density plot for the first plausible value along with basic descriptive statistics, and (*b*) the items, numbered and ordered along the *x*-axis from easiest to most difficult. Vertical dashed lines are the demarcation points for proficiency levels from 1c to 6. Proficiency levels are developed to better contextualize scores. Historically, PISA included 6 proficiency levels with 1 being the lowest and 6 being the highest. However, over time proficiency level 1 was expanded and now includes levels 1a, 1b, and 1c. As such, in math a score of 233.17 to less than or equal to 295.47 falls within PISA's lowest proficiency level, suggesting that students at this level can typically "respond to questions involving easy to understand contexts where all relevant information is clearly given in a simple, familiar format (e.g., a small table or picture) and defined in a very short, syntactically simple text. They are able to follow a clear instruction describing a single step or operation" (OECD 2019c, chap. 15, p. 20). The item locations indicate the point along the PISA scale where examinees are measured. For instance, we can see that examinees meeting benchmark 1c are measured by just two items. And examinees that meet benchmark 2 are measured by the most items. We highlight important aspects of each plot subsequently. The plots for reading and science, which tell a story consistent with the results for math, were created but not included in the current article due to space constraints.

All math results are located in figure 2, in which each country is represented by a panel. The results for Ecuador, which was the top PISA-D country in math, are located in figure 2*A*. Notably here and throughout, the mean for just the first plausible value—also reported in the plot—will be slightly different from the overall mean because of small variations across these random draws. The mathematics mean in Ecuador is 377—more than a full standard deviation below the historic PISA mean of 500. We can see that items measuring proficiency in Ecuador match substantial portions of the distribution. Nevertheless, even in this relatively high-performing country, two or fewer items measure examinees that fall one standard deviation or more below the mean $(377 - 77 = 300)$. This translates to a full 16 percent of examinees, using the usual assumption that 68 percent of normally distributed observations are within one standard deviation of the mean.

In Honduras and Guatemala, in figure 2*B*–*C*, the math mean is 343 and 334 points, respectively. Based on the item locations, students below the means in either country are measured by just six items. Unfortunately, the situation deteriorates with performance. Looking at the lowest-performing country, Zambia (fig. 2*G*), which has a math achievement mean of 258, students around the mean are measured by a single item. This suggests that students below the average level of achievement in Zambia are measured by no items at all.

Next, we present the results for the probability of a correct response from hypothetical representative students to a typical math item in each country, which are located in table 3. In this case, typical refers to the item, described above, that has a response probability close to the historic mean of 500. Here, we can see that the probability of a correct response for an average student to our typical item ranges between .11 in Zambia to .30 in Ecuador. Given that this item is multiple choice with four response options, in all countries except Ecuador, the probability of a correct response is lower than if the student just guessed. For higher-performing students (those that are one standard deviation above their country mean), probabilities range from .21 in Zambia to .48 in Ecuador. For low performers, probabilities range from .06 in Zambia to .17 in Ecuador. Again, these probabilities are worse than chance. Taken together, we can see that average students within a particular country have low probabilities of a correct response. And among the lower-performing students, there

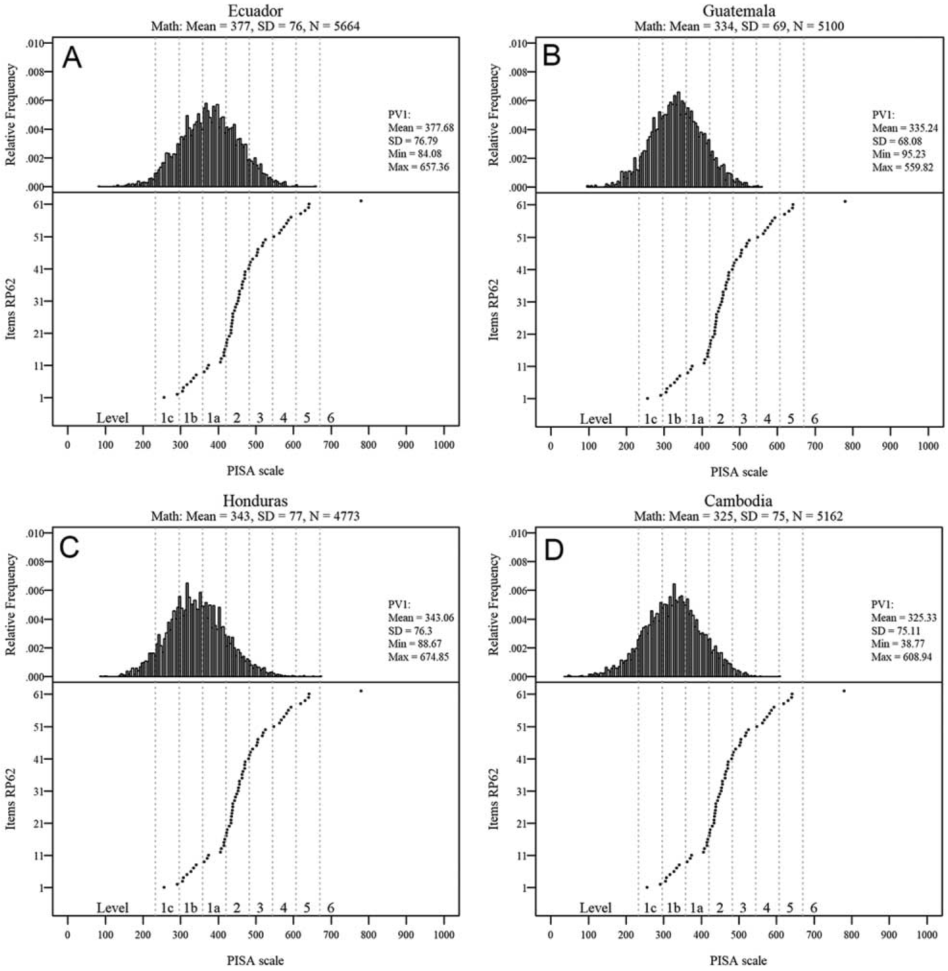

FIG. 2.—Construct maps for mathematics. A color version of this figure is available online.

is very little chance of a correct response to an item of average difficulty. These results show substantial mismatch between the population of examinees in PISA-D and the test intended to measure them. Many students are unmeasured by the test and, in most countries, a student that is average in their country will have little chance of correctly answering a typical PISA-D item.

Assessing Measurement Equivalence in PISA-D

The models used to estimate achievement in PISA-D (and all other international assessments) rely on an assumption that the parameters that characterize items (as in eq. 1) are equivalent across the populations of interest. This idea is best demonstrated visually. In figure 3, we illustrate the item

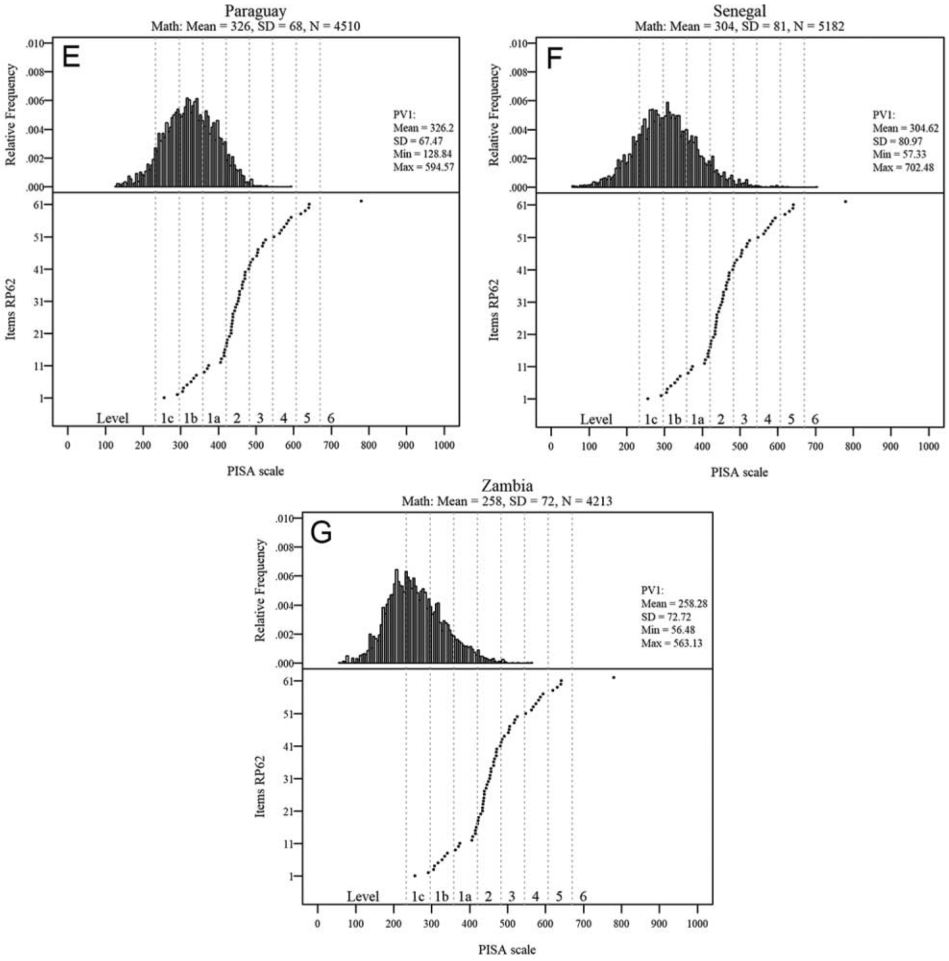

FIG. 2 (*continued*)

response function (IRF) that results from applying equation 1 to the typical PISA-D item, used in the previous section. This is the black curve. In figure 3, the *x*-axis represents the PISA-D proficiency continuum and the *y*-axis represents the probability of a correct answer. Proceeding from left to right across the *x*-axis, we can choose values of proficiency that correspond to probabilities of a correct answer. An especially important point along the *x*-axis corresponds to the difficulty parameter, or b_j, as in equation 1. This corresponds to the location along the proficiency continuum where the probability of a correct answer is .50. Our typical item's difficulty, on the PISA-D scale, is 462.2, or below the historic mean of 500. As noted, we assume that this IRF holds for each population that is administered this item. When this assumption holds,

TABLE 3
PROBABILITY OF CORRECT RESPONSE FOR A TYPICAL MATH ITEM

			Probability of Correct Response for a Student That Is:		
Country	Mean	SD	Below Average	Average	Above Average
Cambodia	325	75	.107	.202	.349
Ecuador	377	76	.166	.299	.477
Guatemala	334	69	.122	.217	.356
Honduras	343	77	.123	.232	.396
Paraguay	326	68	.114	.203	.336
Senegal	304	81	.083	.170	.316
Zambia	258	72	.059	.114	.210

the item is said to be *measurement equivalent*. In other words, for two examinees that have the same level of proficiency, the probability of a correct answer is the same.

In contrast, an item is said to suffer from differential item functioning (DIF), if for two examinees of identical proficiency, the probability of a correct answer is not the same. A visual representation of this is the dashed curve in figure 3. For some population (*population A*), the black curve for our item

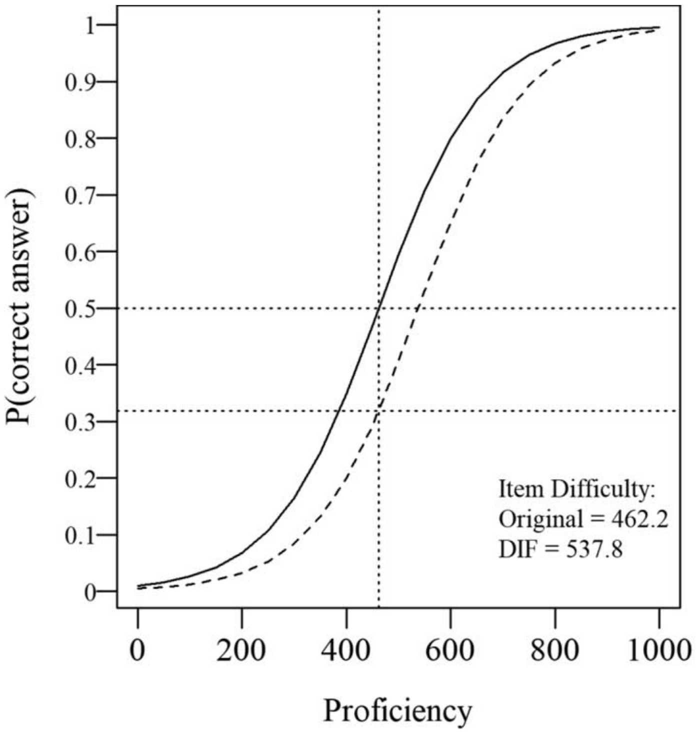

FIG. 3.—Graphical representation of an item with differential item functioning (DIF)

holds; however, for another population (*population B*), the dashed curve holds for the same item. For two examinees of equivalent proficiency—say, exactly at the difficulty of this item, 462.2—but that belong to the two different populations, we can see that the probabilities of a correct answer are drastically different. In *population A*, the probability of a correct answer is .50; however, in *population B*, the probability is .32. Importantly, these two examinees are equally proficient. Intuitively, if an item seems harder (or easier) than it is for a group of examinees, we would wrongly infer that those examinees do not (or do) know the content associated with that item. A consequence is that their score on that item would be lower (higher) than it should be. If DIF is limited to a single item, its impact is limited. When DIF exists for many items, it can have a substantial biasing effect on achievement estimates. As such, it is important to detect DIF in policy-relevant populations. To that end, a number of DIF detection methods exist in the literature.[2] In the case of PISA, the root mean squared deviation (RMSD) (OECD 2017b, 151; Oliveri and von Davier 2011) is used to detect DIF. Setting aside the technical details, the RMSD quantifies the distance between the black and dashed curves in figure 3 for a given country. If the distance exceeds a threshold, then the item is flagged as having DIF. Items with DIF are subjected to a set of standardized procedures defined by the OECD and agreed upon by a panel of experts. These procedures include eliminating the item or allowing the item to have its own parameter. This essentially allows the country in question to use the dashed curve rather than the black curve.

Recent research has uncovered substantial problems with the sensitivity of the RMSD to detect items with positive DIF (an item is harder for the country in question) in low-performing countries (Tijmstra et al. 2019). Again, a visual representation, located in figure 4, is useful to understand why this is the case. When a country is very low performing, as is the case with many PISA-D countries, their empirical IRF—the curve generated by their data—can be similar to the dashed curve in figure 4. Importantly, the curve does not span the entire continuum. Rather, the empirical IRF only covers a small (and lower) portion of the achievement continuum. Because only a small portion of the curve is observed, the RMSD measure only detects a slight departure in the shape of the two curves. This is because the RMSD only measures the distance where both curves exist. At the extreme, it can be shown that an item that is infinitely more difficult for a low-performing country than the overall item difficulty would never be detected by the measure. This fact explains why there are far more DIF items on PISA 2015 in middle-performing countries like the United States than in the lowest-performing country, the Dominican Republic (Tijmstra et al. 2019). The properties of the RMSD indicate that far more items should be identified as having DIF in lower-performing countries than actually are.

[2] See Holland and Thayer (1988); Swaminathan and Rogers (1990); Glas and Jehangir (2014); Svetina and Rutkowski (2014).

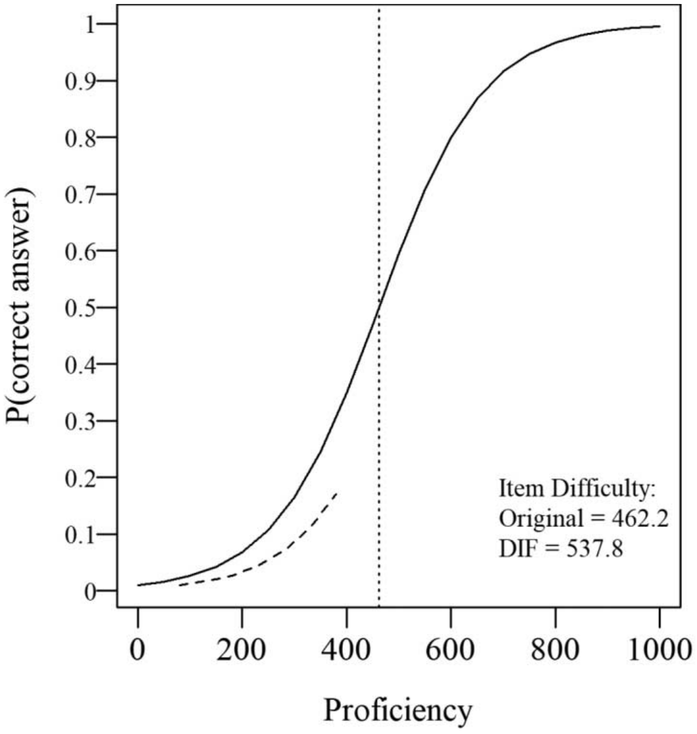

FIG. 4.—Graphical representation of root mean squared deviation insensitivity. DIF = differential item functioning.

Moreover, given that, in low-performing countries, the items will usually have DIF such that the item is more difficult than assumed, this will have an overall penalizing effect on achievement estimates.

Discussion

To appropriately measure students' proficiency, it is imperative that the assessment instrument is tailored to measure the population of interest. In fact, a clear goal of PISA-D was to make the assessment more accessible to low- and middle-performing countries when compared to PISA. However, what our analysis clearly shows is that in many countries PISA-D did not meet the mark. To be clear, for large segments of assessed populations, few or no items measure these students. Even for typical (average achieving) students, the probability of a correct response for an average item is low. This leads us to an important conclusion, first our analysis provides clear and indisputable evidence that currently PISA-D is too difficult for many participating countries. The questions on PISA-D do not align with the proficiency of countries that participated in the assessment. Furthermore, any inferences or interpretations based on PISA-D results are limited, given the validity issues raised by our analysis. It is apparent that the framework is also not well matched to this group of participants because it is simply an expansion of the PISA framework.

Although PISA-D, as an independent study, was a one-off occurrence, the PISA-D instruments will continue as an option for newcomers to PISA in 2021 (OECD 2019b). This makes our findings especially relevant, given that these instruments will be used to measure the batch of new countries, many of whom will be PISA-D veterans or other low-performing countries. In addition, the PISA framework will be expanded, based on the PISA-D framework. Considering our findings, we would argue that this effort risks falling short. To that end, we recommend that the PISA 2021 framework should be further developed and expanded to account for the unique context of low- and middle-income countries if it is to accommodate all participants. Also, much work should be done to understand what a framework should look like if PISA is to measure countries like Cambodia and Norway with the same assessment. This updated framework would have to include a vast expansion of proficiency in all subjects. Of course, with advances in testing technologies this is not unthinkable.

The OECD should do more to develop a well-targeted assessment with links to PISA if it intends to assess proficiency in low-performing systems. In practice, this might involve an altogether different class of items that are suited for participants from economically developing countries that can be linked through a subset of items to the PISA scale. In addition, the OECD would have to pay particular attention to ensure that the introduction of easier items is still in line with the existing framework. As noted, in the current PISA-D, the linking items were PISA trend items, which were much too difficult. Moving forward, the main PISA assessment should introduce more easy trend items so that PISA-D could use those to link the study in later cycles. Although the OECD plans to develop items to enhance measurement at the low end of the continuum for 2021, this innovation is only planned for the computer-based version of PISA (OECD 2019a). Unfortunately, many, if not all, former PISA-D countries are expected to participate in the paper-and-pencil platform, meaning that these countries will not be able to take advantage of specially designed items. In addition, the OECD also needs to focus on better tools for detecting when countries are not being equivalently measured. We have shown that the current tools to ensure proper measurement do not work well for low performers.

Although we point to a number of areas in need of attention on the achievement side of PISA-D, we recognize the OECD for their attention to improving the background questionnaires, which take careful account of the unique and complex context of low- and middle-income countries and students' living and learning situations. The expanded student questionnaire included detailed questions around reasons for extended absences (e.g., absent teacher, pregnancy, need to work, etc.); reasons for being unable to attain educational goals (e.g., could not pay, discrimination, unsupportive family, school is too violent); and an enhanced measure of socioeconomic status, which touched on living situations typically not encountered in economically developed countries. These topics included poverty measures such as home floor composition (dirt, wood, tile, stone, or cement), where they get drinking water (from a personal or shared well, a river, a tanker, piped to home, etc.), whether their parents can read and write, whether the family has a flush toilet, and whether any toilet is shared among people that are not family members. Although improvements to PISA-D background questionnaires are likely warranted, the OECD's limited efforts directed at developing assessment questions stands in contrast to their focus on the background questionnaires. Among other, already stated, reasons, poor measurement on the proficiency side challenges valid interpretations regarding achievement differences for these groups. In other words, the test-examinee mismatch in PISA-D makes it hard to determine whether achievement differs for students from, for example, poor- versus well-resourced homes, regardless of the care with which the background questionnaire was developed.

Then, who has responsibility in such a situation? First, we contend that the OECD needs to be more upfront when assessing low-performing countries. The organization needs to be clear that the test is most likely too difficult for some countries and in those countries the assessment does not provide much information about what the students know and can do. In addition, the OECD needs to be clear that they are measuring a specific type of proficiency and not a universal proficiency. PISA, recall, is a test developed for and by OECD member countries. As a reminder, PISA-D is designed to measure students on the PISA scale. That said, the OECD provides extensive frameworks explaining what they intend to measure for both PISA and PISA-D. In addition, the OECD releases all its assessment data, allowing researchers to dig into such issues. In other words, in this regard, the organization practices a level of transparency not seen in all testing situations.

Our findings have clear policy and research implications. First, considering the item-ability mismatch, PISA-D countries are not ready to participate in main PISA. In fact, participation in PISA would most likely exacerbate the problems outlined in this article and result in even fewer students from PISA-D countries being appropriately measured. Further, our study provides clear evidence that if the OECD is truly concerned with developing a more appropriate assessment for low-performing countries the organization will have to include more questions at the lower end of the proficiency continuum. One obstacle to such a design is that item creation is expensive and such a process would certainly add to the cost of PISA-D. Finally, given the gross item to proficiency mismatch, any research or policy recommendations resulting from PISA-D should be approached with caution. When the data are used authors need to discuss the measurement limitations and the clear threats to valid interpretation of results.

References

Adams, R. J., and J. Cresswell. 2016. "PISA for Development Technical Strand A: Enhancements of PISA Cognitive Instruments." OECD Education Working Paper no. 126, OECD Publishing, Paris.

Addey, C., and S. Sellar. 2018. "Why Do Countries Participate in PISA? Understanding the Role of International Large-Scale Assessments in Global Education Policy." In *Global Education Policy and International Development: New Agendas, Issues and Policies*, ed. Antoni Verger, Hulya K. Altinyelken, and Mario Novelli. London: Bloomsbury.

Auld, E., J. Rappleye, and P. Morris. 2019. "PISA for Development: How the OECD and World Bank Shaped Education Governance Post-2015." *Comparative Education* 55 (2): 197–219. https://doi.org/10.1080/03050068.2018.1538635.

Carr-Hill, R. 2015. "PISA for Development Technical Strand C: Incorporating Out-of-School 15-Year-Olds in the Assessment." OECD Education Working Paper no. 120, OECD Publishing, Paris.

Embretson, S. E., and S. P. Reise. 2000. *Item Response Theory for Psychologists*. Mahwah NJ: Erlbaum.

Engel, L. C., D. Rutkowski, and G. Thompson. 2019. "Toward an International Measure of Global Competence? A Critical Look at the PISA 2018 Framework." *Globalisation, Societies and Education* 17 (2): 117–31. https://doi.org/10.1080/14767724.2019.1642183.

Glas, C., and K. Jehangir. 2014. "Modeling Country-Specific Differential Item Functioning." In *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, ed. L. Rutkowski, M. von Davier, and D. Rutkowski. Boca Raton, FL: CRC Press.

Hambleton, R. K., and H. Swaminathan. 1985. *Item Response Theory: Principles and Applications*. Berlin: Springer.

Holland, W. P., and D. T. Thayer. 1988. "Differential Item Performance and the Mantel-Haenszel Procedure." In *Test validity*, ed. H. Wainer and H. Braun. London: Routledge.

Kirsch, I., J. de Jong, D. Lafontaine, J. McQueen, J. Mendelovits, and C. Monseur. 2002. "Reading for Change: Performance and Engagement across Countries: Results from PISA 2000." OECD Publishing, Washington, DC. https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33690904.pdf.

Lockheed, M., T. Prokic-Bruer, and A. Shadrova. 2015. "The Experience of Middle-Income Countries Participating in PISA 2000–2015 (PISA series)." OECD Publishing, Paris.

Meyer, H.-D., and A. Benavot. 2013. *PISA, Power, and Policy: The Emergence of Global Educational Governance*. Oxford: Symposium.

MoEYS (Ministry of Education, Youth and Sport). 2018. *Education in Cambodia: Findings from Cambodia's Experience in PISA for Development.* Cambodia: MoEYS.

OECD. 2000. "Measuring Student Knowledge and Skills: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy." OECD Publishing, Paris.

OECD. 2010. "PISA 2009 Assessment Framework." OECD Publishing, Paris. https://doi.org/10.1787/9789264062658-en.

OECD. 2012. "PISA 2009 Technical Report." OECD Publishing, Paris. http://www.oecd.org/edu/preschoolandschool/programmeforinternationalstudentassessment pisa/pisa2009technicalreport.htm.

OECD. 2016a. "PISA for Development." OECD Publishing, Paris.

OECD. 2016b. "PISA for Development: Benefits for Participating Countries" (PISA for Development Brief No. 2). OECD Publishing, Paris. https://www.oecd.org/pisa/aboutpisa/PISA-for-Development-Benefits-for-participating-countries-PISA-D-Brief2.pdf.

OECD. 2017a. "PISA 2015 Technical Report." OECD Publishing, Paris. http://www.oecd.org/pisa/data/2015-technical-report/.

OECD. 2017b. "Scaling PISA Data." In PISA 2015 Technical Report. OECD Publishing, Paris.

OECD. 2019a. "PISA 2021 Assessment and Analytical Framework." OECD Publishing, Paris.

OECD. 2019b. "PISA 2021 Integrated Design." OECD Publishing, Paris. https://www.oecd.org/pisa/pisaproducts/PISA2021_IntegratedDesign.pdf.

OECD. 2019c. "PISA for Development 2018 Technical Report." OECD Publishing, Paris. https://www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport/.

OECD. n.d. "PISA Governing Board." OECD Publishing, Paris. https://www.oecd.org/pisa/contacts/pisagoverningboard.htm.

Oliveri, M. E., and M. von Davier. 2011. "Investigation of Model Fit and Score Scale Comparability in International Assessments." *Psychological Test and Assessment Modeling* 53 (3): 315–33.

Rutkowski, D., L. Rutkowski, and Y.-L. Liaw. 2018. "Measuring Widening Proficiency Differences in International Assessments: Are Current Approaches Enough?" *Educational Measurement: Issues and Practice* 37 (4): 40–48. https://doi.org/10.1111/emip.12225.

Rutkowski, L., D. Rutkowski, and Y.-L. Liaw. 2019. "The Existence and Impact of Floor Effects for Low-Performing PISA Participants." *Assessment in Education: Principles, Policy and Practice* 26 (6): 643–64. https://doi.org/10.1080/0969594X.2019.1577219.

Sjøberg, S. 2015. "OECD, PISA, and Globalization: The Influence of the International Assessment Regime." In *Education Policy Perils.* London: Routledge.

Svetina, D., and L. Rutkowski. 2014. "Detecting Differential Item Functioning Using Generalized Logistic Regression in the Context of Large-Scale Assessments." *Large-Scale Assessments in Education* 2 (1): 4. https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-014-0004-5.

Swaminathan, H., and H. J. Rogers. 1990. "Detecting Differential Item Functioning Using Logistic Regression Procedures." *Journal of Educational Measurement* 27 (4): 361–70. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x.

Tijmstra, J., Y.-L. Liaw, M. Bolsinova, L. Rutkowski, and D. J. Rutkowski. 2019. "Sensitivity of the RMSD for Detecting Item-Level Misfit in Low-Performing Countries." *Journal of Education Measurement* 57 (4): 566–83.

von Davier, M., E. Gonzalez, and R. J. Mislevy. 2009. "What Are Plausible Values and Why Are They Useful?" *IERI Monograph Series* 2:9–36.

Wilson, M. 2004. *Constructing Measures: An Item Response Modeling Approach.* London: Routledge.