unesco
Institute for Statistics

GLOBAL ALLIANCE TO MONITOR LEARNING

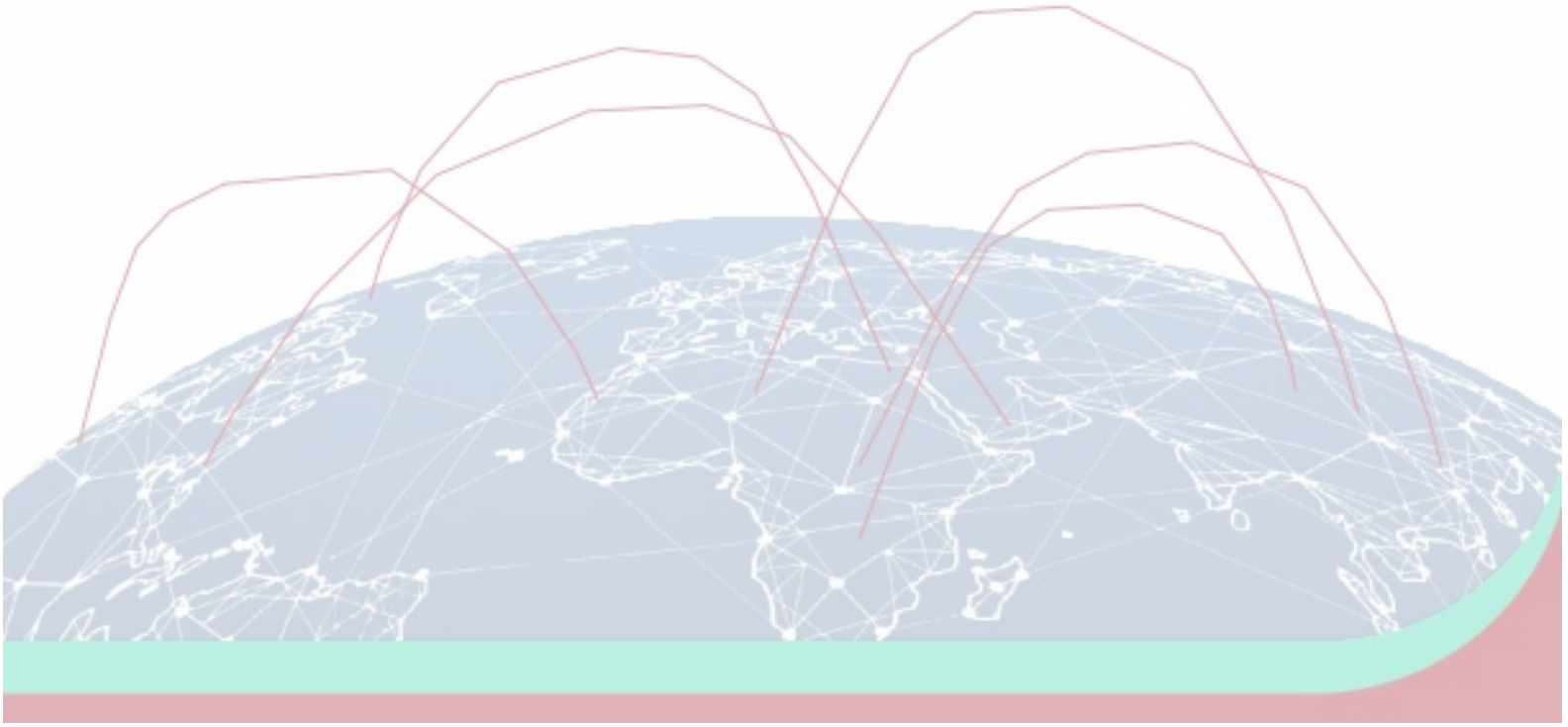
2024 CONFERENCE ON EDUCATION DATA AND STATISTICS

WG/GAML10/CNI/7

The advantages of regional large-scale assessments: Evidence from the ERCE learning survey

Tenth meeting of the Global Alliance to Monitor Learning (GAML)

Paris, 6 - 7 December 2023



International Journal of Educational Development 102 (2023) 102867

Contents lists available at ScienceDirect

International Journal of Educational Development

journal homepage: www.elsevier.com/locate/ijedudev

The advantages of regional large-scale assessments: Evidence from the ERCE learning survey

Diego Carrasco [a,*], David Rutkowski [b], Leslie Rutkowski [b]

[a] Centro de Medición MIDE UC, Escuela de Psicología, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Macul, Santiago, Chile
[b] Counseling & Educational Psychology, Indiana University, 201 N Rose Ave, Bloomington, IN 47405, United States

ARTICLE INFO

Keywords:
International assessment
Examinee-test alignment
ERCE
Wright maps

ABSTRACT

This study examines the potential advantages of regional assessments, such as ERCE 2019, in addressing challenges faced by larger international large scales assessments with heterogeneous populations. The paper investigates whether a regionally focused assessment, developed with the active involvement of all participating countries and targeting more homogeneous populations in terms of language, culture, and economic development, can result in better alignment between measurement instruments and participants' proficiency. Using construct mapping techniques and item response theory reliability indexes, the study aims to identify whether the measurement gaps observed in studies with more heterogeneous populations studies like TIMSS and PISA also exist in ERCE.

1. Introduction

In spite of the care with which International Large Scales Assessments (ILSAs) in education are designed, they remain subject to a cross-cultural measurement paradox: 'The larger the cross-cultural distance between groups, the more likely cross-cultural differences will be observed, but the more likely these differences may be influenced by uncontrolled variables' (Van De Vijver and Matsumoto, 2011, p. 3). In the context of recent ILSA administrations, this issue is difficult to avoid, as more (and more heterogeneous) countries are included in each study. For instance, the Programme for International Student Assessment (PISA), the largest ILSA, began with 44 participating educational systems in 2000. By 2022, the study expanded to include 82 participating educational systems. A notable aspect of PISA is the participation of both Organisation for Economic Co-operation and Development (OECD) member countries and non-member countries, commonly referred as "partner countries". Since all OECD countries participate in every cycle, PISA's growth primarily stems from the inclusion of partner countries, which are generally less economically developed. Another example of ILSA expansion is evident in the Trends in International Mathematics and Science Study (TIMSS). In 2019, TIMSS included 70 systems participating systems in either fourth or eighth grade, which is an increase of 30 systems from the study's first administration in 1995. Like PISA, TIMSS included highly varied systems in regard to economic development. Table 1 highlights a facet of the heterogeneity in PISA by illustrating the growth in the number of participating educational systems and their respective GDP per capita, expressed in 2018 US dollars. Notable is the growth in partner countries as well as the substantial difference in economic development between these two groups of participating educational systems.

Different reasons have been laid out in the literature regarding why countries participate in ILSAs. These reasons include external factors such as regulations, normative reasons related to countries conforming to global accountability practices, and rational reasons linked to public policy making (Ahmed et al., 2022; Liu and Steiner-Khamsi, 2022). For instance, participation in certain ILSAs promoted by the OECD is often expected for member states. However, in the case of Mexico, there was a temporary suspension of its participation in PISA 2022 as the pilot studies were put on hold (El Financiero, 2021). However, following the public announcement of this news in April 2021, Mexico's president reaffirmed the country's participation in the OECD study (Carrillo, 2021). On the other hand, Mexico declined its participation in ERCE 2025. Liu and Steiner-Khamsi (2022) suggest that low- and middle-income countries may be inclined to participate in ILSAs due to normative expectations, following the example of other countries in their region, engage in test-based accountability. Further, participating in ILSAs can serve as indirect tools to attract international donors, loans, and aid, and it can also put countries on the map and facilitate policy

* Correspondence to: Centro de Medición MIDE UC, Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile.
E-mail address: dacarras@uc.cl (D. Carrasco).

https://doi.org/10.1016/j.ijedudev.2023.102867
Received 3 May 2023; Received in revised form 15 July 2023; Accepted 26 July 2023

Table 1
Number of OECD and partner countries since 2000 with GDP per capita.

Year	Number of OECD Countries	Avg GDP per capita (in 2018 USD)	Number of Partner Countries	Avg GDP per capita (in 2018 USD)
2018	35	$37,225	44	$16,541
2015	34	$36,810	37	$15,149
2012	34	$41,819	31	$22,952
2009	34	$40,767	40	$17,856
2006	28	$39,836	27	$16,763
2003	30	$33,354	11	$18,212
2000	28	$27,965	15	$13,556

borrowing and lending among participating educational systems (Wagemaker, 2013). Implementers and national coordinators emphasize the rational motivations for participating in ILSAs, such as evidence-based policy making and decision making in participating countries (Ahmed et al., 2022; Lagos, 2021). Regardless of the specific reasons for countries enrolling in ILSA studies, participation in ILSAs has witnessed growth, particularly in low- and middle-income countries over the past two decades (Rutkowski and Rutkowski, 2019).

Regarding the methodological challenges that arise when assessing heterogeneous populations, three areas often pose difficulties for ILSAs (Rutkowski and Rutkowski, 2019): defining, operationalizing, and measuring comparable constructs; instrument translation; and drawing representative samples. The growth in participation of international assessments has made the assessment development process increasingly complex, leading to potential threats to the valid interpretations and uses of assessment results. Notably, Rutkowski and Rutkowski (2019) discuss the challenges of reaching all populations in all countries, how translation goes beyond being a purely technical pursuit when dealing with 90 different systems, and how the increasingly broad range of achievement necessitates that assessments expand the scope of what and how they measure. Concerning the latter, research has demonstrated that both PISA and TIMSS do not adequately measure all participating countries, particularly among the lowest performing groups (Rutkowski and Rutkowski, 2019; Rutkowski, Rutkowski, and Liaw, 2019).

Both TIMSS and PISA exhibit poor alignment between the difficulty of administered items and average achievement of low-performing populations. This leads to *floor effects*, whereby a large proportion of students in a particular country score at or near the minimum score possible on a test, resulting in a lack of variability in scores, which makes it difficult to accurately compare the performance of students across countries. Such misalignment leaves large areas of the achievement continuum under- or essentially unmeasured (Rutkowski, Rutkowski, and Svetina Valdivia, 2022). Even on an assessment especially tailored to low-performing countries, substantial misalignment exists for many participating countries, with extreme misalignment leaving some populations essentially unmeasured (Rutkowski and Rutkowski, 2021). The primary focus of this paper is to address the issue of test difficulty-student proficiency misalignment and investigate whether a more regionally focused assessment results in better alignment. One potential solution to the challenges mentioned above is the development of cross-cultural assessments that target educational systems with greater homogeneity in language, culture, or geography. Established in 1995, the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) was one of the first such assessments. SACMEQ's policy and programs are determined by the 16 Ministers of Education that comprise the consortium's governing council. Since 1995, the SACMEQ consortium has administered four periodic assessments that measured mathematics and reading. Unlike TIMSS and particularly PISA, SACMEQ is designed and implemented by local experts who develop and include measures of important region-specific topics. For instance, in 2007, an HIV and AIDS survey and assessment was introduced to measure student attitudes and knowledge in this critical area. Murimba (2005) argued that the regional focus of SACMEQ's assessment and background questionnaires resulted in better tailored and more meaningful measurement when compared to larger ILSAs.

Another regional assessment system, developed by the UNESCO Regional Bureau for Education in Latin America and the Caribbean (LLECE), was established in 1994 by the 15 founding members with aims to: promote evidence-based education policy through the generation of (empirical) data on quality education and associated factors; develop education assessment capacities; and serve as a forum to generate and share ideas and discuss best practices in education (Viteri and Zoido, 2019). To date, the LLECE laboratory has completed four regional assessments with the most recent being the fourth ERCE, completed in 2021.

Building on the primary focus mentioned earlier, the purpose of the current paper is to examine whether regional assessments with more homogenous populations, such as ERCE 2019, can better attend to some of the measurement challenges faced by larger ILSAs that feature more heterogenous populations. ERCE 2019 assessed students' academic abilities at third and sixth grades in the domains of language, mathematics, and science, with the latter only assessed in sixth grade. We chose the ERCE 2019 study due to its desirable features that may minimize some challenges of the larger ILSA studies noted by Rutkowski and Rutkowski (2019) and because it represents the most recent published cycle of this study. In the following sections, we describe each of these features in relation to ERCE 2019 characteristics.

In 2019, ERCE included 15 participating countries in Latin America and was administered in only two languages: Spanish with a small number of language adaptations (UNESCO, 2022 see Annex 2 for adaptations), and Brasilian Portuguese. This greatly reduced the language variability when compared to other ILSAs. By comparison, the TIMSS 2019 assessment was translated into 50 languages (Martin et al., 2020) and PISA had over 100 country-specific language combinations (OECD, 2020a). It is also important to note that the source language for questions in TIMSS is usually English, which is then translated into the other languages. In the case of ERCE, the test and questionnaires language are either only Spanish or Brasilian Portuguese, further reducing the burden of translation and possible interpretation errors or cultural differences in understanding.

Regarding economic differences, the range in ERCE was much less drastic than TIMSS and PISA. For example, the wealthiest country in ERCE was Uruguay (17,278 USD, GDP per capita 2018) and the poorest was Nicaragua (2021 USD, GDP per capita 2018). By comparison, in PISA the economic differences are much more drastic. For example, the Dominican Republic (7947 USD, GDP per capita 2018) who participated in PISA 2019 was compared in math, science, and reading to Ireland with a GDP per-capita of 99,152 USD. Given more similar economic situations across participating countries, ERCE was able to include background questionnaires that are more aligned across countries, resulting in greater comparability compared to PISA and TIMSS (Sandoval-Hernández, Miranda, Rutkowski, and Matta, 2018; UNESCO, 2022). In regard to curriculum, the ERCE 2019 achievement portion of the study was developed based on a curricular study covering all participating countries (UNESCO, and LLECE, 2020; UNESCO-OREALC, 2020; Vanni and Valenzuela, 2020). Furthermore, all participating countries contributed to item development for each test, and importantly, representatives from all countries participated in discussions around the inclusion of content domains and coverage to ensure a well-balanced assessment for each participating country (UNESCO-OREALC, 2016). This practice contrasts significantly with PISA 2018, for example, where only 17 and 14 countries contributed items to the computer- and paper-based assessments, respectively, with Serbia being the only non-OECD member contributing items (OECD, 2020b). Additionally, only OECD members and associate members have voting rights on the development of PISA, excluding the majority of PISA participating educational systems (OECD, n.d.).

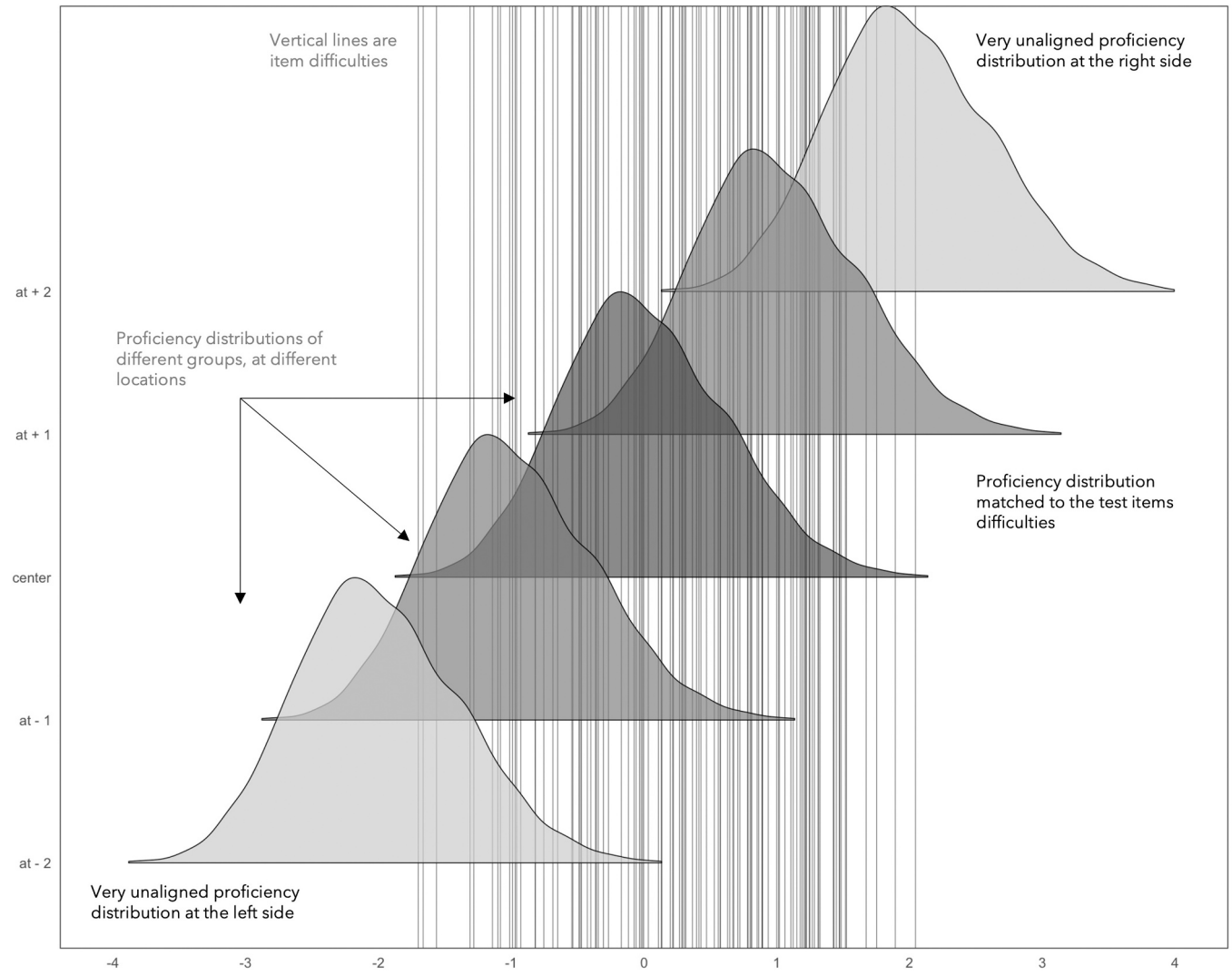


Fig. 1. Diagram for item location and country proficiency distributions.

In summary, ERCE's design emphasizes measuring the curricula of a relatively more homogeneous group of countries in terms of language, culture, and economic development compared to TIMSS and PISA. Moreover, unlike other larger international studies, ERCE was developed with the active involvement of all participating countries and its framework was informed by a comprehensive curricular study. Finally, through a deliberative development process, ERCE countries had equal opportunities to influence study implementation decisions via the "national coordinators assembly" (Vanni and Valenzuela, 2020). With these characteristics in mind, we investigate whether a regionally focused assessment can result in better alignment between the measurement instruments and the student's proficiency distribution of each participating educational system. To do so, we use construct mapping techniques (Wilson, 2005) used in previous research (see Rutkowski and Rutkowski, 2021; Rutkowski et al., 2019), and item response theory reliability indexes, to query whether the same sorts of measurement gaps observed in studies like TIMSS and PISA exist in ERCE. We describe the data and our analytic approach subsequently.

2. Methods

2.1. Data

We use data from the UNESCO's Fourth Regional Comparative and Explanatory Study, ERCE 2019 (UNESCO, 2022). This is an international large scale assessment study, that collects representative samples of third and sixth grade students, from 16 Latin-American countries. Students are assessed in math, language, and science (the latter only at sixth grade). Further, context and background information from students, teachers, school principals, and student's families is collected. This study uses a two-stage sampling design. Schools are selected through a stratified design, and all students from the sampled classroom from the selected school are invited to participate.

2.2. Measures

Most of the ERCE items are multiple choice (95% on average across tests), where student responses can be classified into correct and incorrect answers. A small number of open-ended questions are included in the math and science test, eliciting short answers scored as correct, partially correct, or incorrect. The specification of each test considered a number of subject domains, and the involvement of three cognitive processes for each test. For example, the assessment of math at sixth grade includes the domains of numeracy, geometry, measurement, statistics, variation, patterns and algebra. Moreover, the test is design to also assess the cognitive processes of recognition of objects and elements, simple problem solving, and complex problem solving (UNESCO-OREALC, 2023). Students from each participating country answer a

Table 2
Sample Size, Achievement scores and estimated Latent means, and number of items above and below countries Latent Means (sixth grade students, Mathematics).

	GDP		ERCE scores		Latent Means		EAP-PSR reliability
Countries		n	E	CI95%	E	CI95%	E
República Dominicana	8314	4899	636	[630; 642]	-0.48	[− 0.55; − 0.43]	0.70
Panamá	15,069	5632	645	[639; 650]	-0.41	[− 0.46; − 0.37]	0.73
Paraguay	5774	4849	647	[641; 654]	-0.38	[− 0.45; − 0.34]	0.75
Guatemala	4254	4895	657	[650; 664]	-0.30	[− 0.38; − 0.25]	0.75
Nicaragua	1983	4868	663	[658; 668]	-0.26	[− 0.31; − 0.22]	0.69
El Salvador	4003	5920	676	[671; 681]	-0.15	[− 0.22; − 0.11]	0.73
Honduras	2499	4423	682	[672; 693]	-0.11	[− 0.22; − 0.02]	0.75
Cuba	8027	5126	689	[679; 699]	-0.05	[− 0.16; 0.04]	0.83
Argentina	12,712	5004	690	[684; 696]	-0.04	[− 0.11; 0.02]	0.77
Colombia	6390	4467	707	[699; 714]	0.09	[0.00; 0.16]	0.78
Ecuador	5854	6758	720	[712; 728]	0.20	[0.12; 0.27]	0.80
Costa Rica	10,170	3699	726	[719; 733]	0.25	[0.17; 0.32]	0.75
Brasil	8638	4349	733	[724; 742]	0.31	[0.21; 0.39]	0.82
México	9820	4824	758	[752; 764]	0.51	[0.43; 0.57]	0.80
Perú	6611	5938	759	[751; 767]	0.52	[0.45; 0.58]	0.81
Uruguay	16,036	5176	759	[753; 766]	0.53	[0.47; 0.58]	0.82

Note: GDP = gross domestic product in US dollars, with constant prices at 2010, source https://datos.bancomundial.org/; n = sample size; E = estimated mean; Ci95% = lower and upper limit of the 95% confidence interval; ERCE scores = ability scores using the metric of the ERCE study, with an expected mean of 700 and standard deviation of 100 scores; Latent means = estimated ability means using the fitted IRT model; EAP-PSR reliability = expected a posteriori person separation reliability of the person latent mean realizations per country.

reduced number of items, generated with a rotated block design with two blocks of 16 questions. Thus, a student answers about 32 items, while the specific items depend on the randomly assigned form. On the full ERCE assessment 35%, 19%, 23%, 13%, and 10% of items assess numeracy, geometry, measurement, statistics and patterns and algebra, respectively. While in terms of cognitive process 17% of the items involved the cognitive processes of recognition of objects and elements, 32% measured simple problem solving, and 32% measured complex problem solving (UNESCO-OREALC, 2023). The full study administration requires two days, where students complete the test in the first day and, on the second day, they complete the background questionnaire. Students are allocated 60 min to complete each subject-area test, with the exception of the math test in sixth grade, which is allocated 70 min.

2.3. Analysis

For our analysis, we rely on item response theory (IRT), which is a latent variable model that relates the probability of a correct answer to parameters that describe the item and the examinee. In particular, we use a Rasch model (Rasch, 1960) that relates person proficiency and an item's location to the probability of a correct response, given as:

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

where θ_i is the proficiency level for examinee i. The parameter b_j is a characteristic of item j. In particular, 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% chance of a correct answer. Higher values indicate a more difficult item, lower values represent easier items. For example, an item that is located at the overall average of 0 is more difficult than an item that is located at − 0.25 and is less difficult than an item that is located at 0.25. In a similar vein, examinees that are located at 0 are said to be more proficient than examinees at − 0.25 and less proficient than examinees at 0.25 on the scale. A comprehensive description of these parameters and their interpretations are well outside the scope of this manuscript; however, interested readers are encouraged to consult Hambleton et al. (1991) or Embretson and Reise (2000).

Following a similar approach to Rutkowski et al. (2019), and Rutkowski and Rutkowski (2021), we examine the degree to which ERCE is well-matched to the participating populations of students using a visual means to relate examinees proficiency to items parameters, referred to as a construct map (Wilson, 2005). 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. To plot the proficiency distribution for each country, we use the latent realizations of a multiple group IRT model, which we describe subsequently. These graphical representations give an overall picture of the alignment between a group of examinees to the test. A test that is well-matched to 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. Fig. 1 depicts several hypothetical countries that vary in terms of their proficiency alignment to a hypothetical test. Each curve represents the proficiency distribution of a hypothetical country. Each vertical line represents the location of a hypothetical item along the proficiency/difficulty continuum. Those groups on the far left and far right of the figure are countries with poor proficiency-difficulty alignment. In particular, countries at the left of the figure would be poorly measured by this hypothetical test because the items are overly difficult. The opposite is true for countries at the right of the figure – this test is overly easy for these groups of examinees. Countries in the middle of the figure have proficiency that is well aligned to the test.

To estimate item parameters and population achievement distributions, we use the model in Eq. 1 fit to each group simultaneously as a multi-group IRT model (Millsap, 2011). For our construct maps, we use the following parametrization: we fixed the average of the country latent means 0 (Embretson and Reise, 2000) and within this constraint, we freely estimate relative latent means and variances. Item parameters were constrained to be equal across countries. We used a logit link, and a robust maximum likelihood estimator in the Mplus software (Asparouhov and Muthén, 2020; Muthén and Muthén, 2017). The resulting item parameters and population achievement estimates were combined into construct maps which show the degree to which the examinee proficiency is measured in each population and where any gaps in measurement might exist.

To describe how well examinees from each country are matched to the test, we use two strategies. The first is a visual summary where we display the items' location of the test overlayed as a kind of curtain on each country's proficiency distribution, similar to Fig. 1. This latter strategy helps to summarize how well the test is aligned to the

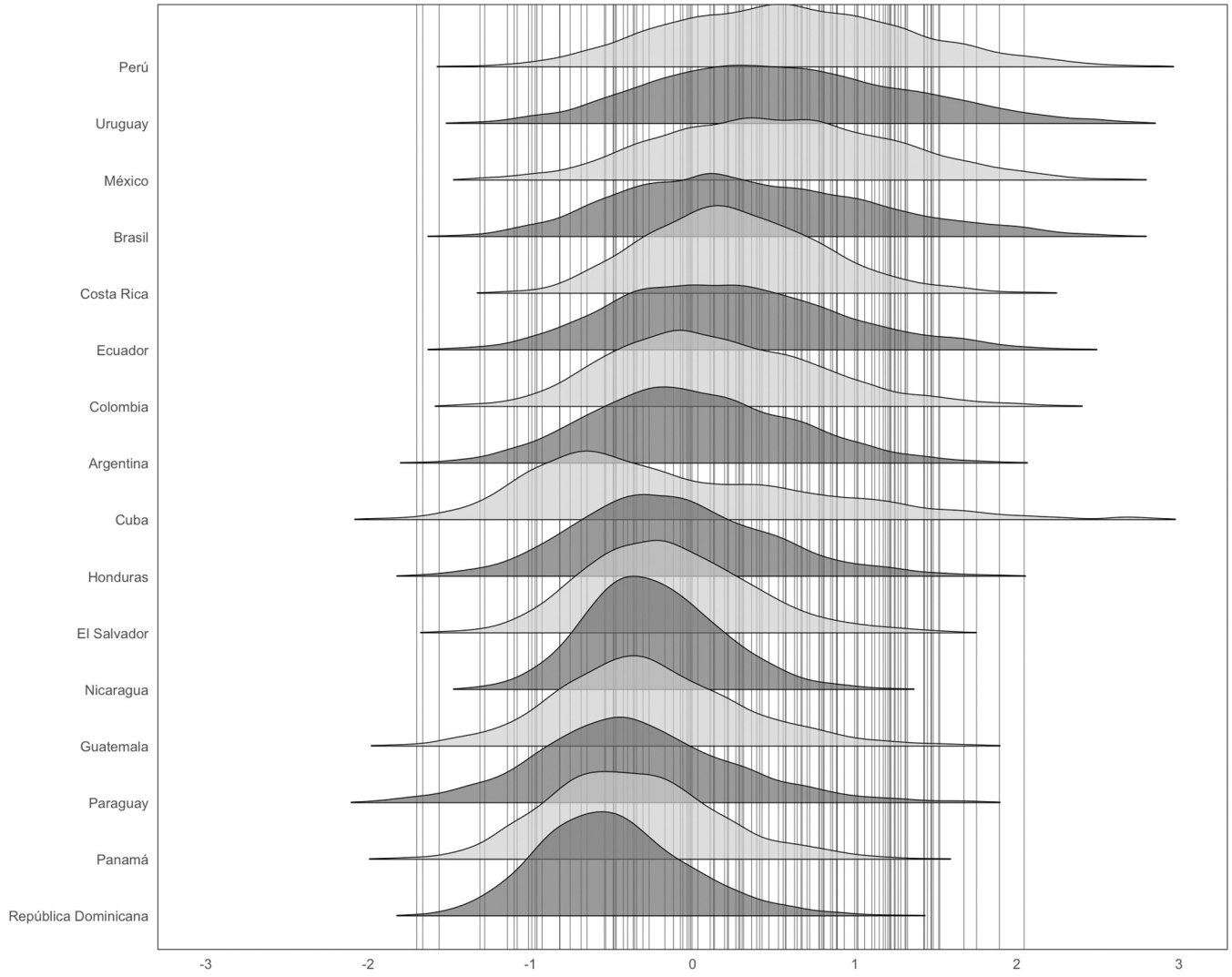


Fig. 2. Countries distribution of the person estimates, for sixth graders Math test, and item location estimates.

participating countries of the study. The second approach we used, was to estimate the expected a posteriori person separation reliability estimate (EAP-PSR) (Adams, 2005). Person separation reliability estimates are the proportion of variance accounted for by person proficiency location, in comparison to the total variance of the measurement process. If a country, or a portion of the distribution of persons, is unaligned to its test, then the standard errors around the person's expected locations will be larger. As such, the more unalignment there is between a group and the test items difficulty distributions, the smaller should be the EAP-PSR reliability estimates.

In summary, we expect that countries with a poor alignment to the distribution of difficulties of the test's items, will display lower reliability (EAP-PSR). In addition, unaligned countries will be located further away from the item location estimates, displayed as an overlayed curtain as described previously. Unaligned countries, similar to the hypothetical countries at the extremes at the right side, or the left side from Fig. 1, should depart from the range of difficulties of the test. In a floor effect scenario, the unaligned countries should be further the left. The opposite should be true in a country with poor alignment because of an overly easy test. Thus, in this latter scenario the test would show ceiling effects for this participating country. This conjecture is consistent with Rutkowski et al. (2021) who showed that in a version of PISA for developing countries (PISA-D) with a collection of easier items, there are considerable floor effects. For example, students from Guatemala and Honduras, two countries participating in ERCE and in PISA-D, feature just 6 items below their country mean in PISA-D, leaving large segments of the proficiency continuum under- or unmeasured. A more extreme case was Zambia where no items existed below their proficiency mean, suggesting that there was little to learn about what their students knew and could do.

3. Results

3.1. Country latent means and reliability (EAP-PSR)

In Table 2, we summarize the ERCE 2019 participating countries' results, including their gross domestic product per capita (GDP), the student sample size, the observed score on the test, the estimated latent mean, and estimated EAP-PSR reliability. Given that latent means across countries average to zero, we can evaluate how far a country is from *average* by noting how far above or below zero their achievement estimate is. Countries exhibited latent mean distances from the center of the proficiency distribution ranging from − 0.48 to 0.53. The EAP-PSR reliability estimates vary from 0.70 at the lowest to 0.82 at the maximum, while the EAP-PSR reliability is 0.80 for the pooled sample of countries.

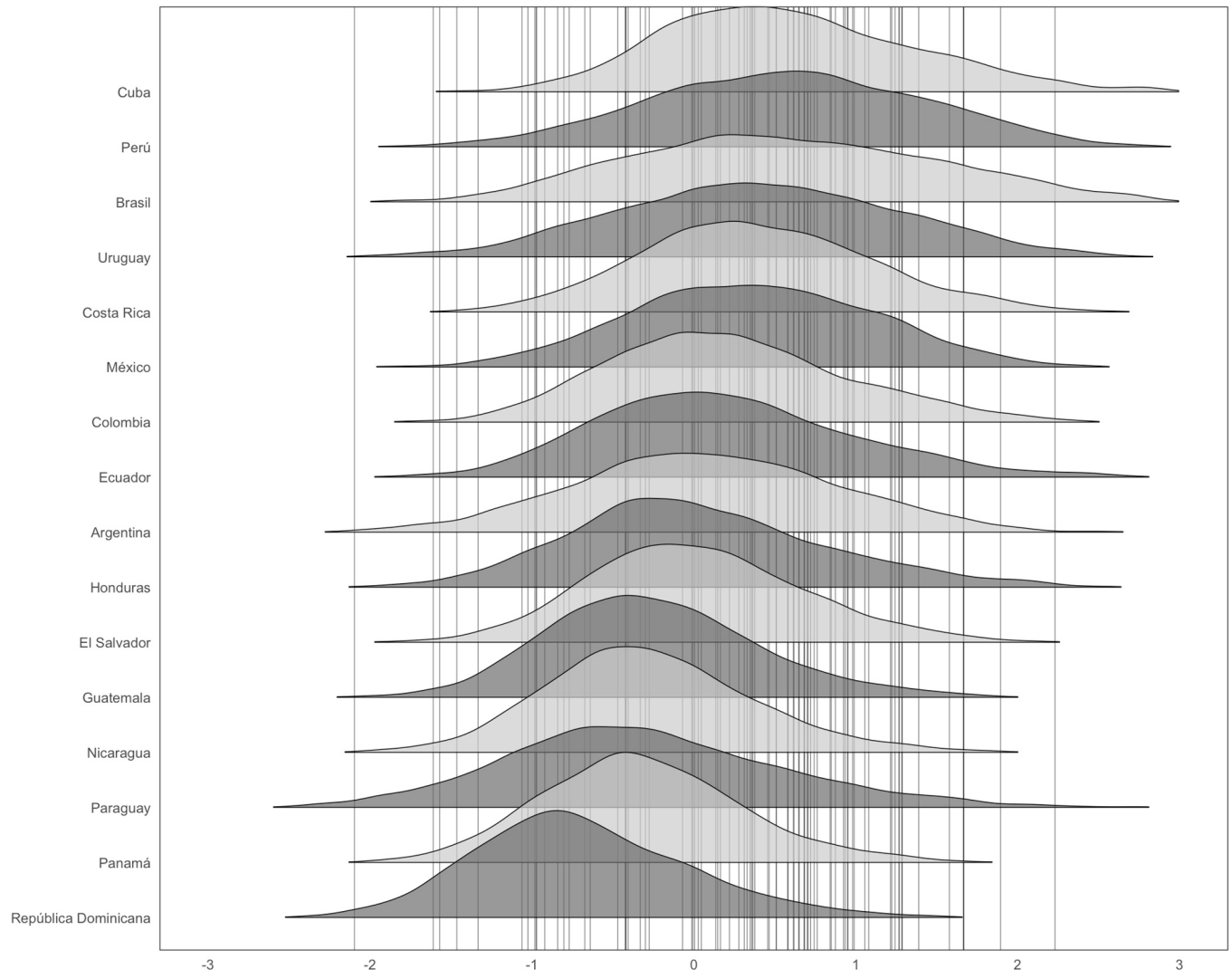


Fig. 3. Countries distribution of the person estimates, for third graders Math test, and item location estimates.

3.2. Distribution of item locations and country proficiency distributions

A limitation of global indicators, such as the EAP-PSE reliability, is that this index cannot tell where the information gaps are. A country can be unaligned at the left, at the right, or the test can have considerable gaps of non-observed items at portions of the ability distribution. Fig. 2 complements the previous results by providing the specific location of all the items of the test along the proficiency continuum for each country. Based on these results, all participating countries seem reasonably well covered by the distribution of item locations of the test. There is no evidence of substantial misalignment for any of the ERCE participating countries, nor is there evidence of floor or ceiling effects. We also include results for math at third grade in Fig. 3, and corresponding figures for the rest of the domains and grades are included Appendix A.

4. Discussion

To measure adequately and with sufficient precision, it is important that an assessment is aligned with the measured populations' proficiency. As noted previously, recent research found that ILSAs that are designed to measure dozens of heterogeneous populations suffer in this regard. In fact, even in PISA-D – a study designed specifically to measure economically developing and lower-performing countries – there were few or no items to measure large segments of proficiency for participating countries. A possible explanation for these measurement gaps is that no new items were developed for PISA-D (OECD). Instead, existing PISA items and items pulled from other extant assessments were used to measure study participants. And for larger studies like TIMSS and PISA, the sheer number and heterogeneity of participating educational systems make it challenging to measure well with a single, common instrument.

For example, translating PISA into over 100 country-specific language combinations (OECD, 2020a) and the substantial difference in achievement across participating educational systems are just a few of the complexities that testing organizations face when measuring a large collection of educational systems. In contrast, a regional assessment like ERCE measures a more homogeneous group of participating countries who generally speak a common language with only slight local variations, with the exemption of Brazil, sharing a Spanish or Portuguese colonial history, and other cultural similarities. The homogeneity of language across the participating countries reduces the burden of translation and local variations, which is a challenge to cross-cultural research (van de Vijver, Jude, and Kuger, 2019).

It is possible that curricula of ERCE participating countries is more similar to each other than to countries outside the region. Most of the ERCE participating countries present a "problem resolution" approach for the teaching of mathematics (UNESCO-OREALC, 2020). We believe

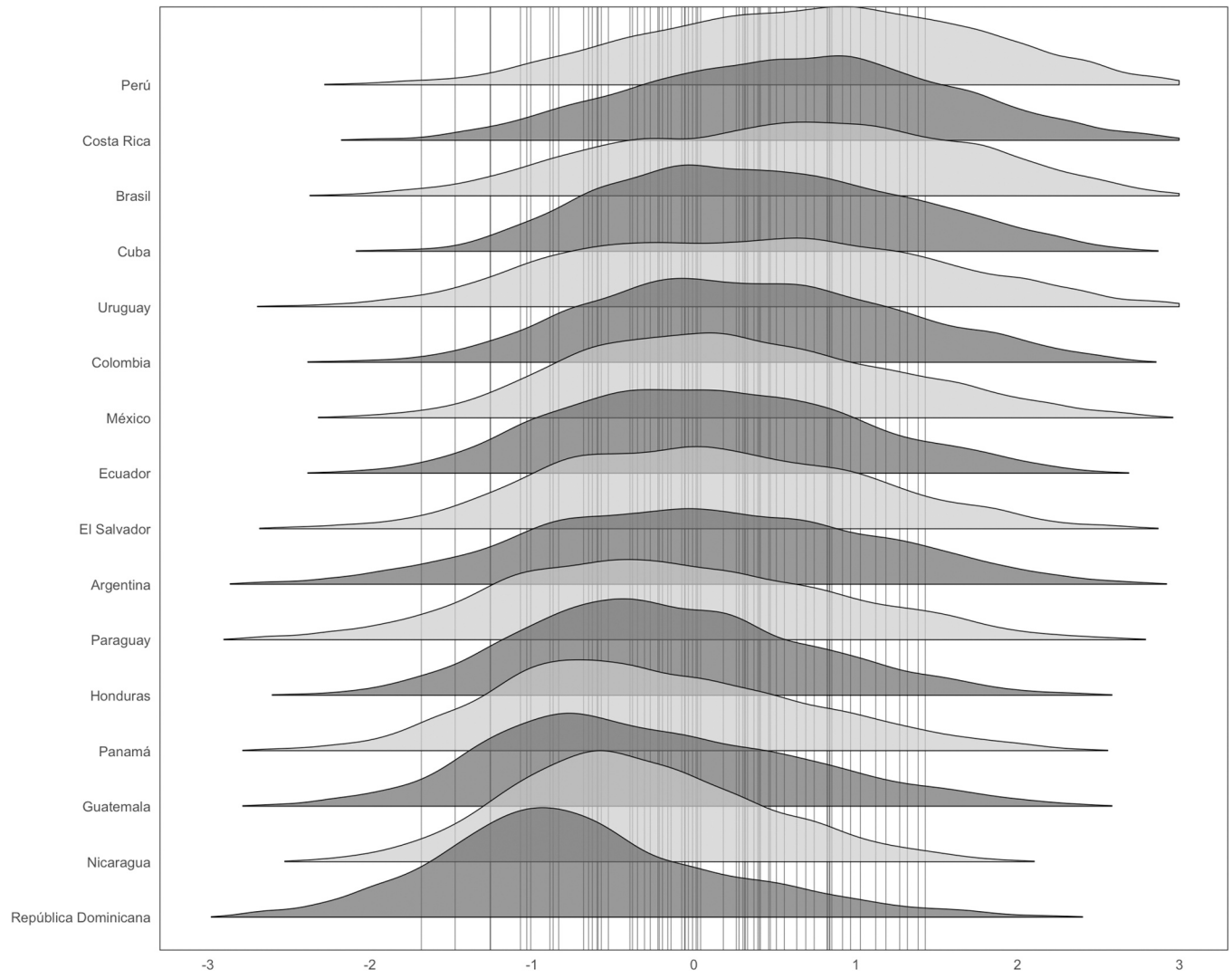


Fig. A1. Countries distribution of the person estimates, for third graders Language test, and item location estimates.

that the active participation by all participating countries in test development, as a collaborative endeavor (Vanni and Valenzuela, 2020), may help reach a higher alignment between country proficiency distributions and item locations in ERCE. Moreover, test development is preceded by a curricular study of all participating countries, thus assuring test selected items for the pilot stage have enough proficiency coverage for each participating country (UNESCO, and LLECE, 2020; UNESCO-OREALC, 2020; Vanni and Valenzuela, 2020). This last feature of the ERCE study could be of particular interest for educators with a focus on research inquiries that require curricular alignment, that the ERCE study provides while other ILSAs may lack.

Our findings provide evidence that in at least one example, a regional approach to cross national assessment confers important advantages when compared to larger international assessments. Although comparison and learning from others was an original goal of international assessment architects (Purves, 1987), massive growth led a number of critics to question the utility of comparing vastly different systems that have little in common (Meyer and Benavot, 2013; Sellar and Lingard, 2013; Sellar, Thompson, and Rutkowski, 2017). For example, comparing Iceland, a small homogenous Nordic country of approximately 350,000 people to Mexico with a diverse economy and a population of nearly 130 million is difficult at best and misleading at worst. Although, admittedly, all systems within a region have important differences, they are often more similar to one another than to systems outside the region, which makes policy borrowing more reasonable as regional peers could be grounded in a similar reality. Thus, the present features of ERCE highlighted in the present study, should be of interest for policy makers, when choosing what study to use for evidence-based policy making, and decision making as ILSAs' studies differed on the quality of information these can provide at different proficiency levels. Further, studies with other regional large-scale assessment can test the present assumption. For the ERCE study, our analysis provides some evidence that there are tangible measurement advantages of a regional large-scale study.

When examining the measurement of socio-economic status between international assessments, Sandoval-Hernandez et al. (2018) found that although neither TIMSS, PISA, nor ERCE in its third round (also known as TERCE) had acceptable measurement properties, the results from ERCE showed the most promise in terms of comparability. The authors wrote, "as a regional assessment that focuses on similar language groups, cultures, and economies (when compared to PISA and TIMSS), with more focus ERCE should be able to design and administer questionnaires that are better tailored to a specific population" (p. 55). Their results showed that ERCE 2013 was able to develop a socioeconomic background scale that exhibited a higher level of comparability than other international assessments. Further, both PISA and TIMSS included participating systems in their scales that did not meet basic quality standards. One explanation is that these educational systems differed so much in terms of the actual construct of socioeconomic status that the

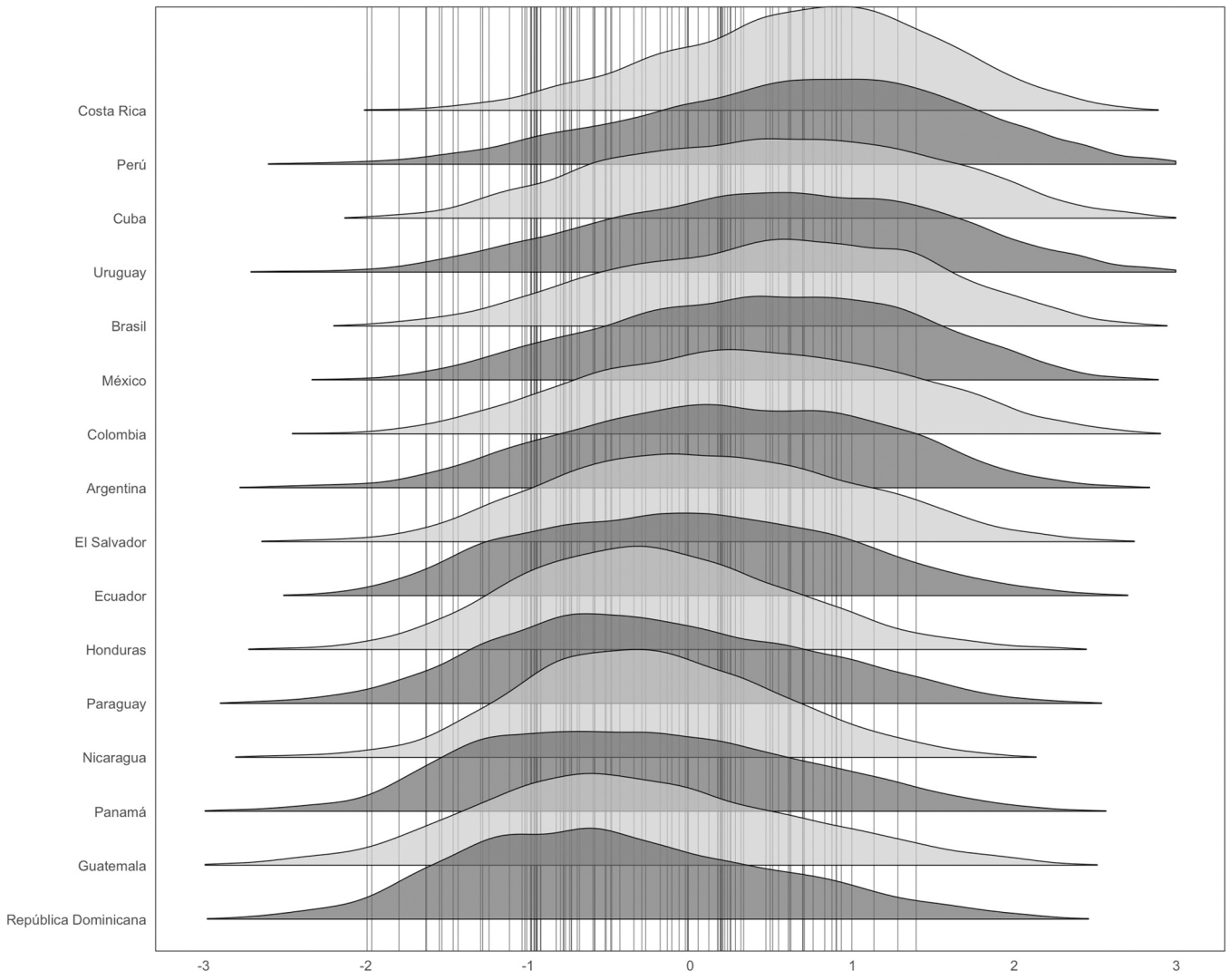


Fig. A2. Countries distribution of the person estimates, for sixth graders Language test, and item location estimates.

developed scale was not relevant. By contrast, ERCE, with a more homogenous set of participating countries, included indicators that were better aligned to the participating set. Sandoval-Hernandez et al.'s findings are consistent with the current paper in demonstrating the advantage of regional assessments on both the achievement and student background side of the assessment.

Although regional assessments provide important benefits to participating countries, larger international assessments offer value beyond what regional assessments can provide. Notably, capacity building in educational systems that have no national or jurisdiction assessment of their own is one clear advantage. Capacity building was an early aim of the first large-scale assessments (Wagemaker, 2013) and remains an important goal of the organizations that conduct the assessments (OECD; Ward, 2019). Clearly regional assessments also offer the possibility of capacity building, however, the resources, institutional knowledge, and infrastructure of large studies like PISA and TIMSS are unrivaled. Second, comparison over time on an internationally agreed-upon measure is valuable. Even if a country is so outside of the reference group of participating countries, internal benchmarking is useful for understanding whether and how an educational system has changed over time on a common international scale. Thirdly, a common limitation of regional studies is their less frequent utilization of data for research inquiries and education policy debates compared to ILSAs such as PISA, which dominates in this aspect. The majority of research in comparative education, policy borrowing, and ILSA studies heavily relies on secondary data from PISA, with over 1000 publications dedicated to it (Hernández-Torrano and Courtney, 2021). In contrast, regional studies often have limited usage due to their documentation primarily being available in the most commonly spoken language of their region of origin. This limitation restricts access to a broader audience of researchers who predominantly use English as the lingua franca. For instance, the documentation for the ERCE study is exclusively in Spanish, while the Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC), another regional study, provides documentation exclusively in French.

Given that regional and international large-scale assessments offer their own distinct advantages (and disadvantages), efforts to assess minimum proficiency levels of academic attainment at the country level should consider what is best to achieve such a goal. The efforts to obtain the proportion of students per country satisfying the United Nations' Sustainable Development Goals 4, Target 4.1.1 (SDG 4.1.1), that is, proportion of girls and boys reaching at least the minimal level of competence in readings, and mathematics, requires assessment test which are informative enough at the expected level of proficiency. Otherwise, percentages of students satisfying such a goal, will include high uncertainty. Our findings show that is not enough to just participate in an international large scale assessment study to report on the SDG 4.1.1, but to participate in the most informative one.

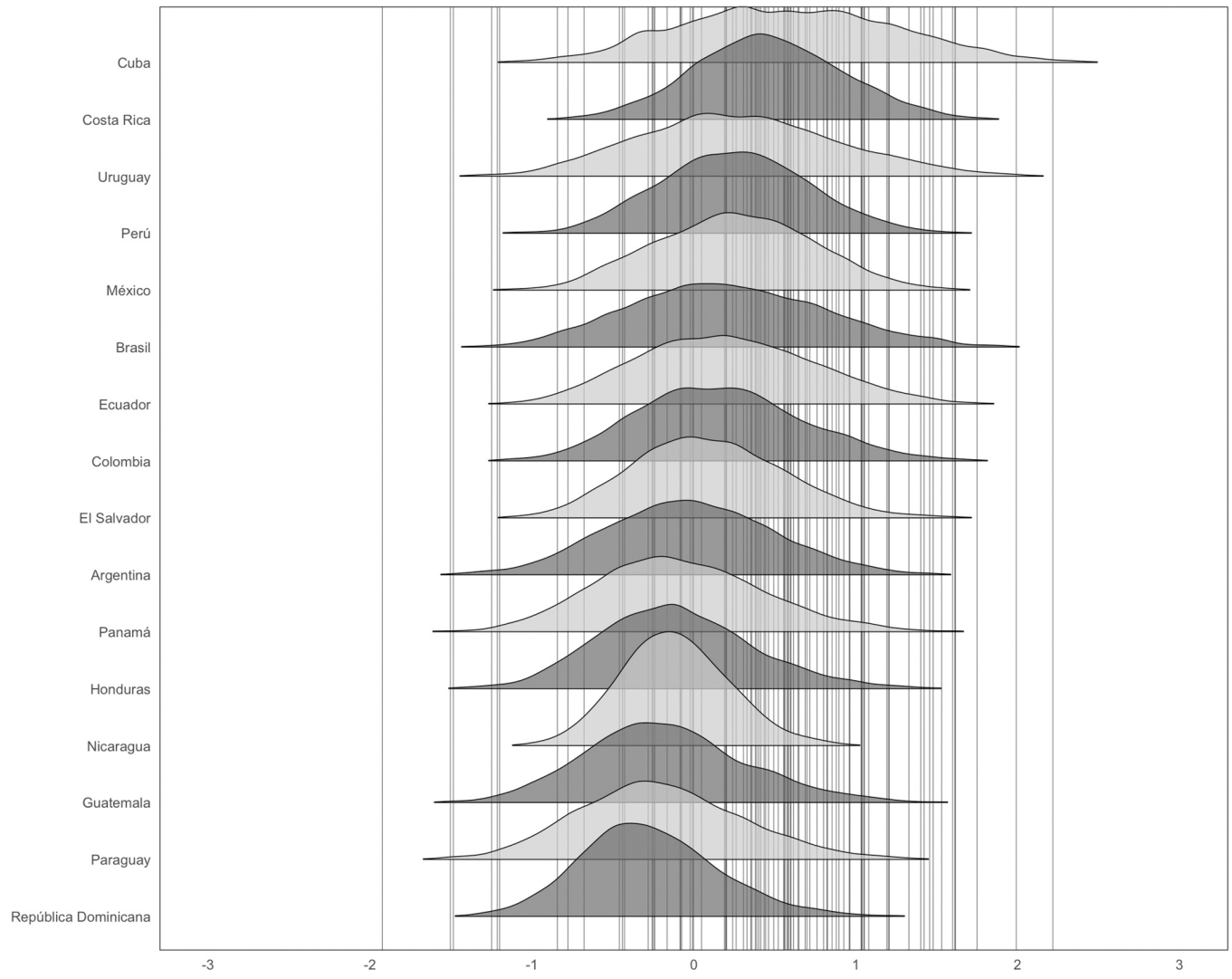


Fig. A3. Countries distribution of the person estimates, for sixth graders Science test, and item location estimates.

A main limitation of the present study is that its claims are limited to the specific features present in ERCE 2019, the regional study being examined, which follows the common research design of ILSAs (e.g., representative sample of students, proficiency scores presented as plausible values, inclusion of background questionnaires with various actors) (see Rutkowski, Gonzalez, Joncas, and von Davier, 2010). According to Lockheed and Wagemaker (2013), ILSA studies should fulfill two missions that need to be aligned: providing measurement quality regarding the outputs of education systems and providing information for policy use, as well as facilitating the generation of such information by participating countries. The present study focuses on ERCE 2019 as a regional ILSA, primarily addressing the measurement quality mission while not fully examining its capacity-building features for policy use. Therefore, the claims made in the present study may not be generalizable to all regional studies that can be considered counterparts of other ILSAs in different countries and regions. Further research is necessary to determine if test difficulty-country proficiency alignment is a common feature in all regional studies or if it is a particular characteristic of the ERCE study analyzed in this study. We believe that test alignment, in this sense, is not achieved solely by setting up a regional study but rather through the results of the test construction process, where test design ensures score precision across the proficiency distribution of the participating populations. Furthermore, the capacity-building aspect and policy use of the present regional study require further research to assess how and to what extent they fulfill the second mission mentioned by Lockheed and Wagemaker (2013).

Participating countries, along with their policy makers and country officials, should have a comprehensive understanding of what they can gain from participating in an ILSA study, taking into account their specific policy interests. For instance, if the primary interest lies in assessing the general competence of students at the end of compulsory secondary education, PISA would be well-suited for the task due to its design, which targets the desired population and provides relevant measures. However, if the primary interest is to obtain population-level results on math, language, and science attainment that are tied to the national curriculum for third and sixth grade, the ERCE study would be a better fit for their purpose. We do not believe it is appropriate to make an "either/or" judgment regarding countries' participation in ILSAs, suggesting that they should only participate in a regional large-scale assessment or an ILSA. Instead, we emphasize that participating countries should be able to assess whether their participation in an ILSA study aligns adequately with their intended objectives. The present study highlights various features of the ERCE study that may have been overlooked but are of general interest to policy makers and educators who require tests aligned to the national curriculum and corresponding to the proficiency distribution of their student population.

5. Conclusion

ERCE is not a perfect measure of student proficiency; however, our findings provide important evidence that ERCE does a better job of ensuring that students are being adequately measured. We believe that curricular-based international large assessment studies face difficult challenges when economic, cultural, or language diversity and large curricular differences between countries are present. The ERCE regional study features a set of participating countries with low language diversity, many cultural commonalities, and economic differences that are not as vast as in other ILSAs. Further, there are meaningful curricular commonalities among participating countries. Moreover, we think the ERCE study implementation takes advantage of the low language barrier and the participation of countries in the test development. Thus, country commonalities and study governance help the study implementation to produce tests that are well-aligned to their participating countries.

CRediT authorship contribution statement

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The corresponding author works as an academic researcher in the research centre (Centro de Medicion MIDE UC, at Pontificia Universidad Católica de Chile) that developed the test and background questionnaires for the Fourth Regional Comparative and Explanatory Study, ERCE, commissioned by UNESCO-OREALC; and is at the present implementing its fifth version.

Appendix

See appendix Figs. A1-A3.

References

Adams, R.J., 2005. Reliability as a measurement design effect. Stud. Educ. Eval. 31 (2–3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008.

Ahmed, S.K., Belisle, M., Cassity, E., Friedman, T., Lietz, P., Spink, J., 2022. Role Int. Large-Scale Assess. (ILSAs) Econ. Dev. Ctries. 119–141. https://doi.org/10.1007/978-3-030-88178-8_7.

Asparouhov, T., Muthén, B. , 2020. IRT in Mplus. In Mplus Technical Appendix. https://www.statmodel.com/download/MplusIRT.pdf.

Carrillo, E. , 2021, May 3). AMLO descarta dejar de aplicar prueba Pisa en México. Forbes México. https://www.forbes.com.mx/amlo-descarta-dejar-de-aplicar-prueba-pisa-en-mexico/.

van de Vijver, F.J.R., Jude, N., Kuger, S., 2019. Challenges in International Large-Scale Educational Surveys. The SAGE Handbook of Comparative Studies in Education. SAGE Publications Ltd,, pp. 83–99. https://doi.org/10.4135/9781526470379.n6.

El Financiero, 2021, May 3. México sí va a seguir en la prueba PISA, asegura López Obrador. El Financiero, 1–8. https://www.elfinanciero.com.mx/nacional/2021/05/03/mexico-si-va-a-seguir-en-la-prueba-pisa-asegura-lopez-obrador/.

Embretson, S., Reise, S., 2000. Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Inc.

Hambleton, R.K., Swaminathan, H., Rogers, D.J., 1991. Fundamentals of Item Response Theory. SAGE Publications.

Hernández-Torrano, D., Courtney, M.G.R., 2021. Modern international large-scale assessment in education: an integrative review and mapping of the literature. Large-Scale Assess. Educ. 9 (1), 1–33. https://doi.org/10.1186/s40536-021-00109-1.

Lagos, E., 2021. Chile: The Challenge of Providing Relevant Information from ILSA Studies for the Improvement of Educational Quality. Improving a Country's Education. PISA 2018 Results in 10 Countries. Springer International Publishing,, pp. 49–82. https://doi.org/10.1007/978-3-030-59031-4_3.

Liu, J., Steiner-Khamsi, G., 2022. Reasons Particip. Int. Large-Scale Assess. 55–73. https://doi.org/10.1007/978-3-030-88178-8_5.

Lockheed, M.E., Wagemaker, H., 2013. International large-scale assessments: thermometers, whips or useful policy tools. Res. Comp. Int. Educ. 8 (3), 296–306. https://doi.org/10.2304/rcie.2013.8.3.296.

Martin, M.O., Davier, M. Von, Mullis, I.V.S. , 2020. Methods and Procedures: TIMSS 2019 Technical Report (M. O. Martin, M. Von Davier, & I. V. S. Mullis, Eds.). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Meyer, H.-D., Benavot, A., 2013. PISA, Power, and Policy: the emergence of global educational governance. Symposium Books.

Millsap, R.E., 2011. Statistical Approaches to Measurement Invariance. Routledge.

Murimba, S., 2005. The Impact of the Southern and Eastern Africa Consortium for Monitoring Educational Quality (Sacmeq. Prospects 35 (1), 91–108. https://doi.org/10.1007/s11125-005-6822-z.

Muthén, L.K., Muthén, B.O., 2017. Mplus User's Guide. Muthén & Muthén.OECD. (n.d.-a). About. https://www.oecd.org/pisa/aboutpisa/ OECD. (n.d.-b). PISA for Development. About. OECD. OECD. (n.d.-c). PISA-D In-School Assessment Technical Report, 8th ed...,. OECD Publishing,. ⟨https://www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport/⟩.

OECD, 2020b. Translation and Verification of the Survey Material. In PISA 2018 Technical report. OECD. https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018 TecReport-Ch-05-Translation.pdf.

OECD , 2020a. PISA 2018 Technical report. In Pisa. OECD. https://www.oecd.org/pisa/data/pisa2018technicalreport/.

Purves, A.C., 1987. The evolution of the IEA: A memoir. Comp. Educ. Rev. 31 (1), 10–28. https://doi.org/10.1086/446653.

Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Nielsen & Lydiche.

Rutkowski, D., Rutkowski, L., 2021. Running the wrong race? the case of pisa for development. Comp. Educ. Rev. 65 (1), 147–165. https://doi.org/10.1086/712409.

Rutkowski, L., Rutkowski, D., 2019. Methodological challenges to measuring heterogeneous populations internationally. The SAGE Handbook of Comparative Studies in Education. SAGE Publications Ltd,, pp. 126–140. ⟨https://sk.sagepub.com/reference/sage-handbook-of-comparative-studies-in-education/i1259.xml⟩.

Rutkowski, L., Gonzalez, E., Joncas, M., von Davier, M., 2010. International large-scale assessment data: issues in secondary analysis and reporting. Educ. Res. 39 (2), 142–151. https://doi.org/10.3102/0013189×10363170.

Rutkowski, L., Rutkowski, D., Liaw, Y.L., 2019. The existence and impact of floor effects for low-performing PISA participants. Assess. Educ.: Princ., Policy Pract. 26 (6), 643–664. https://doi.org/10.1080/0969594X.2019.1577219.

Rutkowski, L., Rutkowski, D., Svetina Valdivia, D., 2022. Multistage Test Design Considerations in International Large-Scale Assessments of Educational Achievement. In: Nilsen, T., Stancel-Piątak, A., Gustafsson, J.E. (Eds.), International Handbook of Comparative Large-Scale Studies in Education. Springer International Handbooks of Education, pp. 749–767. https://doi.org/10.1007/978-3-030-88178-8_63.

Sandoval-Hernández, A., Miranda, D., Rutkowski, D., Matta, T., 2018. Back to the drawing board: Can we compare background scales? Rev. De. Educ. 383, 37–62. ⟨https://dialnet.unirioja.es/servlet/articulo?codigo=6761049⟩.

Sellar, S., Lingard, B., 2013. Looking East: Shanghai, PISA 2009 and the reconstitution of reference societies in the global education policy field. Comp. Educ. 49 (4), 464–485. https://doi.org/10.1080/03050068.2013.770943.

Sellar, S., Thompson, G., Rutkowski, D. , 2017. The Global Education Race. Taking the measure of PISA and International Testing. Brush Education Inc.

UNESCO, 2022. Manual de uso de las bases de datos Estudio Regional Comparativo y Explicativo (ERCE 2019) (Issue Erce). https://unesdoc.unesco.org/ark:/48223/pf0000382518.

UNESCO, & LLECE, 2020 Análisis curricular del ERCE 2019 del conjunto de países que conforman la CECC/SICA. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000375368.

UNESCO-OREALC, 2016. Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo. TERCE,.

UNESCO-OREALC , 2020. ¿Qué se espera que aprendan los estudiantes de América Latina y el Caribe? Análisis curricular del Estudio Regional Comparativo y Explicativo (ERCE 2019). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000373982.

UNESCO-OREALC, 2023. Reporte técnico. Cuarto Estudio Regional Comparativo y Explicativo (ERCE 2019).

Van De Vijver, F.J.R., Matsumoto, D., 2011. Introduction to the methodological issues associated with cross-cultural research. In: Matsumoto, D., van de Vijver, F.J.R. (Eds.), Cross-Cultural Research Methods in Psychology, 1st ed...,. Cambridge University Press, pp. 1–14.

Vanni, X., Valenzuela, J.P. , 2020. Evaluación del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación. https://unesdoc.unesco.org/ark:/48223/pf0000374760.

Viteri, A., Zoido, P., 2019. Building Learning Assessment Systems in Latin America and the Caribbean. The SAGE Handbook of Comparative Studies in Education. SAGE Publications Ltd.,, pp. 600–618. https://doi.org/10.4135/9781526470379.n33.

Wagemaker, H., 2013. International Large-Scale Assessments: From Research to Policy. In: Rutkowski, L., von Davier, M., Rutkowski, D. (Eds.), Handbook of International large-scale assessment: background, technical issues, and methods of data analysis. Chapman and Hall/CRC, pp. 11–36.

Ward, M. , 2019. Capacity Building and Peer Learning in PISA-D and PISA (Issue September). https://www.oecd.org/pisa/pisa-for-development/14._PISA-D_Seminar_2019_Ward.pdf.

M. Wilson Constructing Measures: An Item Response Modeling Approach 2005 Lawrence Erlbaum Associates Publishers.