The advantages of regional large-scale assessments

Evidence from the ERCE learning survey

10 th meeting of the
Global Alliance to Monitor Learning (GAML)
Paris

December 6th
2023

Carrasco, Diego, PhD,
Centro de Medición MIDE UC
Pontificia Universidad Católica de Chile

Rutkowski, David, PhD,
Counseling & Educational Psychology
Indiana University, US

Rutkowski, Leslie, PhD,
Counseling & Educational Psychology
Indiana University, US

Centro UC
Medición - MIDE

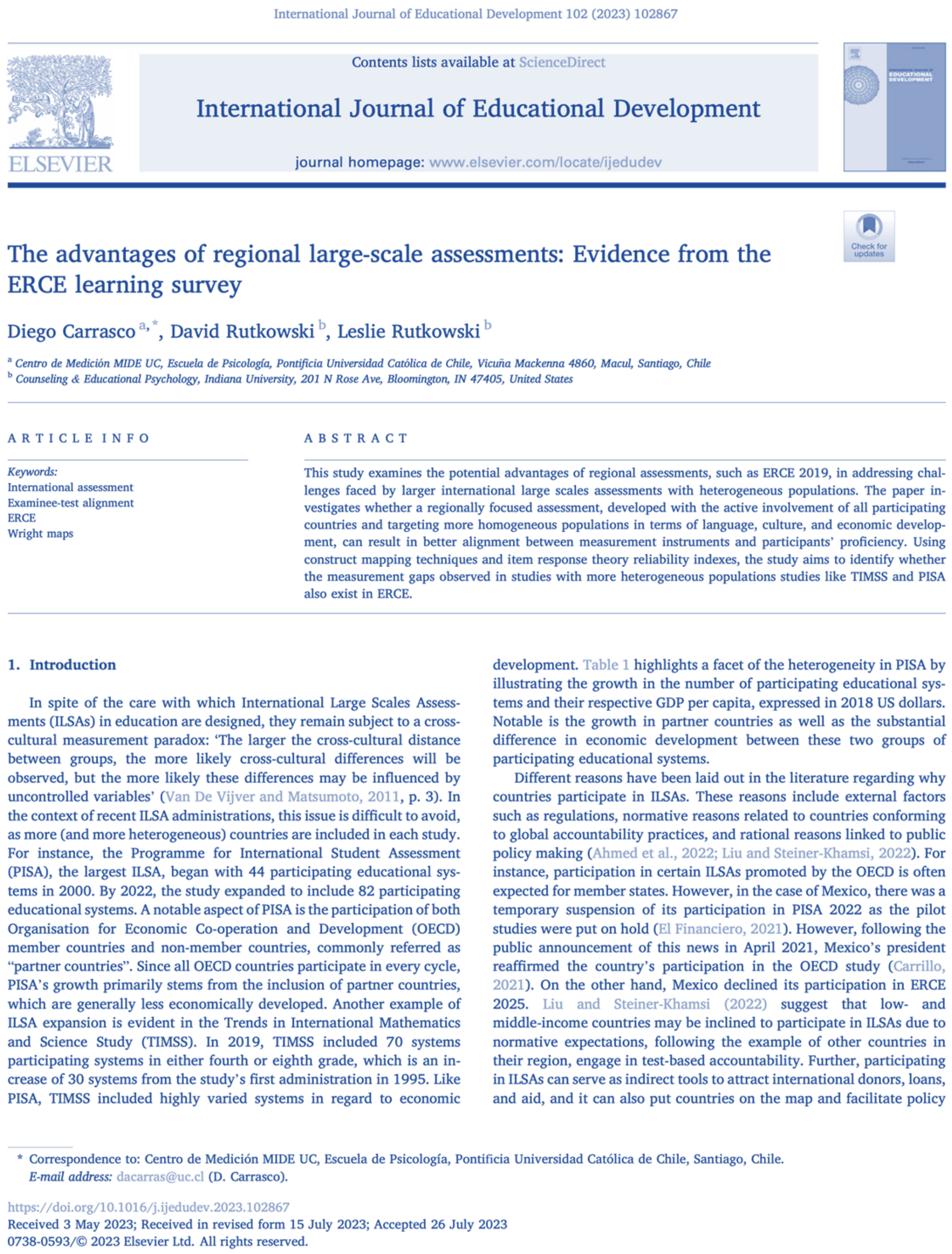
Challenges in ILSA

- It is hard to assess learning for many different educational systems. Several features for large scale assessment are challenging to design comparable and informative test:
- Several **languages**
- **Curricular** differences
- Heterogenous **Ability**

- The higher the **heterogeneity** among the target population of students in a large-scale assessment, the more difficult is to assemble an informative test for all target populations.

- Different ILSA studies present **floor effects** and have difficulties to provide informative results for countries far off from the the test mean difficulty (e.g., Rutkowski et al., 2019; 2022).

- **Regional** large-scale assessment are assumed to **minimize some of these challenges** (e.g., less language diversity, hopefully common curriculum). In the present study we inquiry if ERCE, a regional-ILSA deals better with test-ability alignment.



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

Main problem

Floor effects in ILSA

Uncertainty of scores at different scale locations

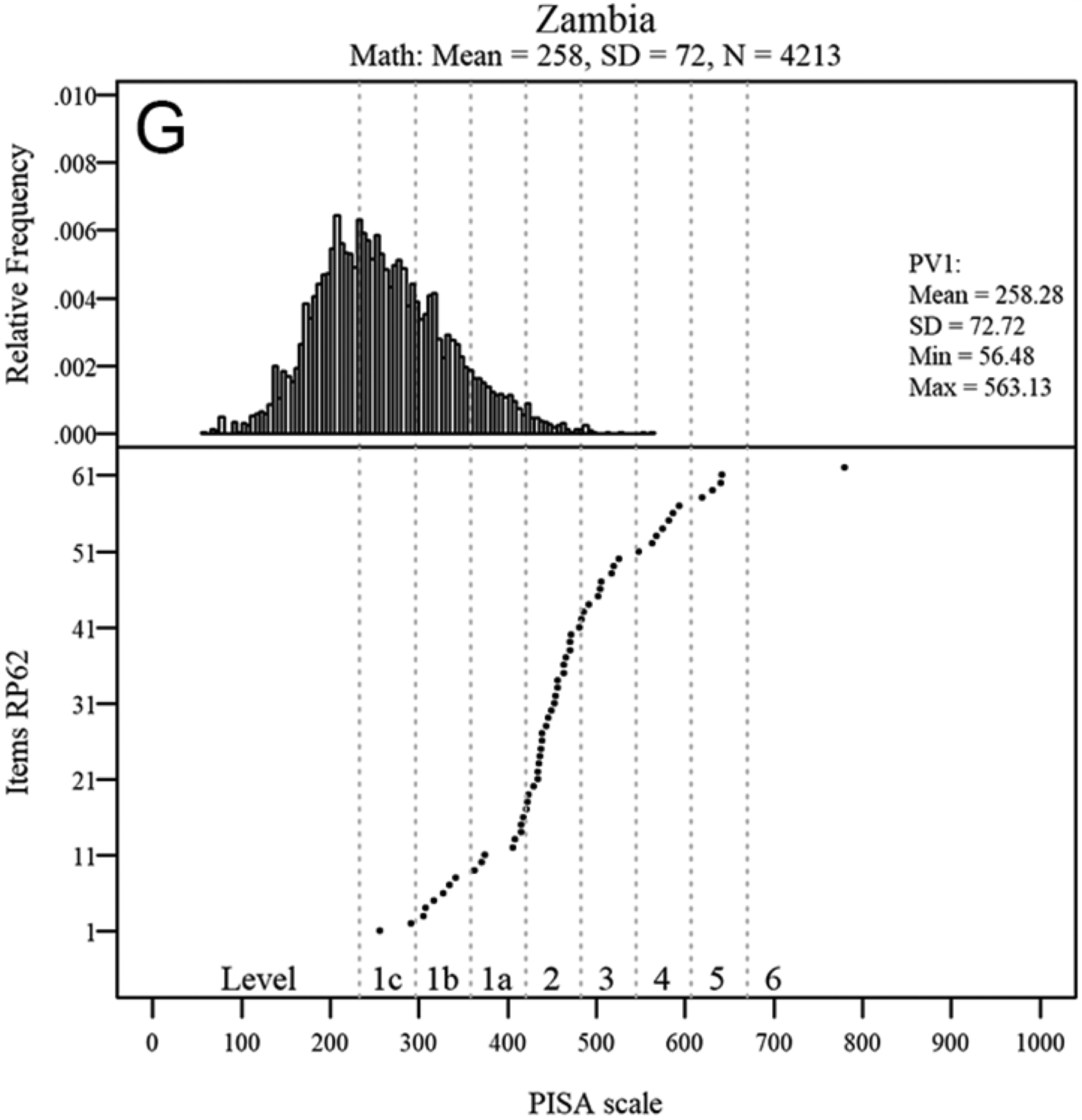
Floor effects

- The further away is a participating country from a test mean difficulty the less reliable are its scores.

- When the ability of a target population is lower than the assemble test, we can expect **floor effects**. These are scenarios when the test won't be as informative for the the lower ability group, because we lack items in said location.

- An extreme case of a floor effect was found for Zambia in PISA-D, where there were no items below the Zambian students means (Rutkowski et al., 2021).

- Similar results have been found in PISA, and TIMSS (Rutkowski and Rutkowski, 2019; Rutkowski, Rutkowski, and Liaw, 2019).



Rutkowski, L., Rutkowski, D., & Liaw, Y. L. (2019). The existence and impact of floor effects for low-performing PISA participants. *Assessment in Education: Principles, Policy and Practice*, *26*(6), 643–664. https://doi.org/10.1080/0969594X.2019.1577219

Floor effects

- The further away is a participating country from a test mean difficulty the less reliable are its scores.

- When the ability of a target population is lower than the assemble test, we can expect **floor effects**. These are scenarios when the test won't be as informative for the the lower ability group, because we lack items in said location.

- An extreme case of a floor effect was found for Zambia in PISA-D, where there were no items below the Zambian students means (Rutkowski et al., 2021).

- Similar results have been found in PISA, and TIMSS (Rutkowski and Rutkowski, 2019; Rutkowski, Rutkowski, and Liaw, 2019).

Rutkowski, L., Rutkowski, D., & Liaw, Y. L. (2019). The existence and impact of floor effects for low-performing PISA participants. *Assessment in Education: Principles, Policy and Practice*, *26*(6), 643–664. https://doi.org/10.1080/0969594X.2019.1577219

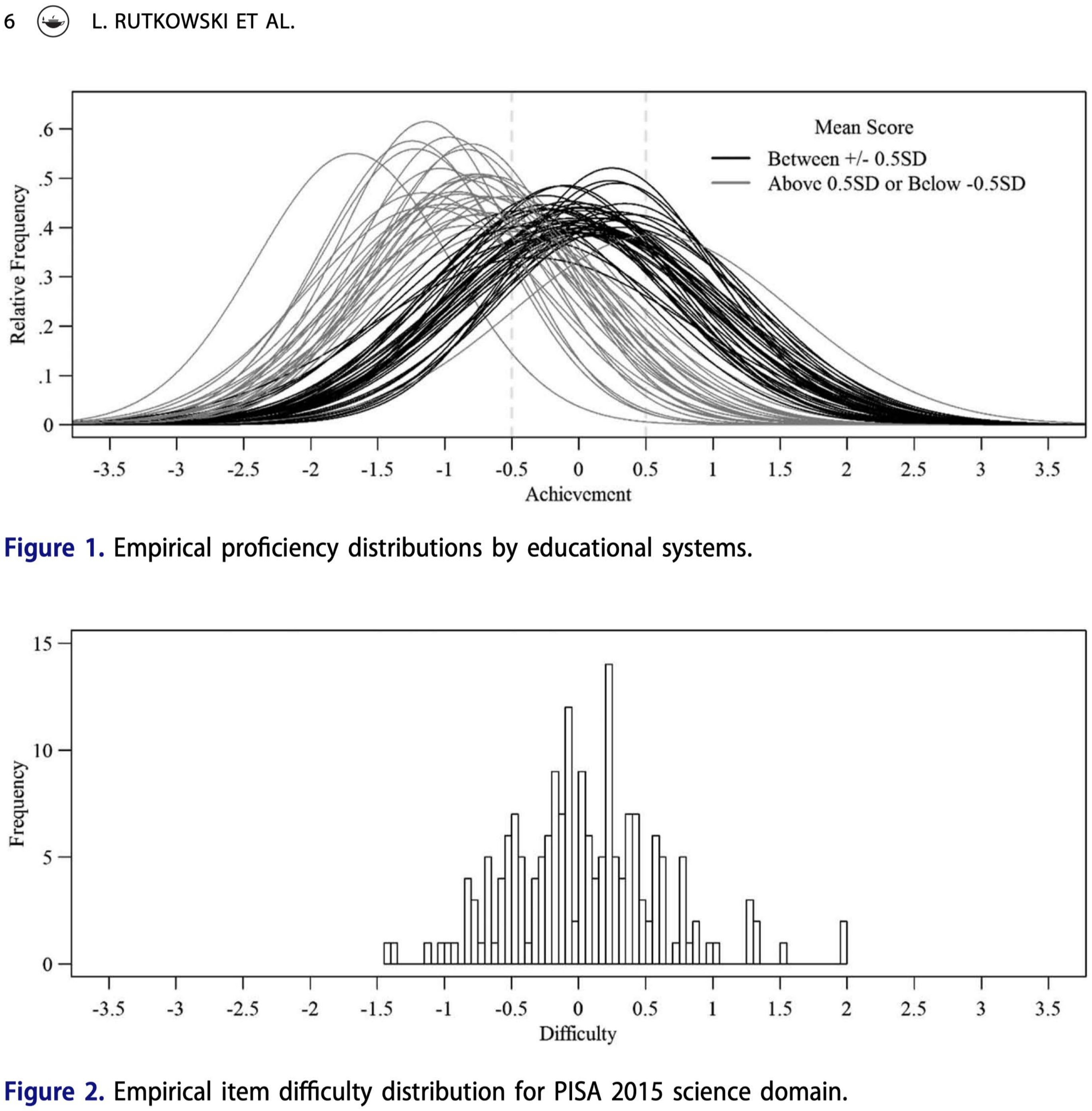

6 L. RUTKOWSKI ET AL.

Figure 1. Empirical proficiency distributions by educational systems.

Figure 2. Empirical item difficulty distribution for PISA 2015 science domain.

Rationale

Test and Ability (miss)match

Uncertainty of scores at different scale locations

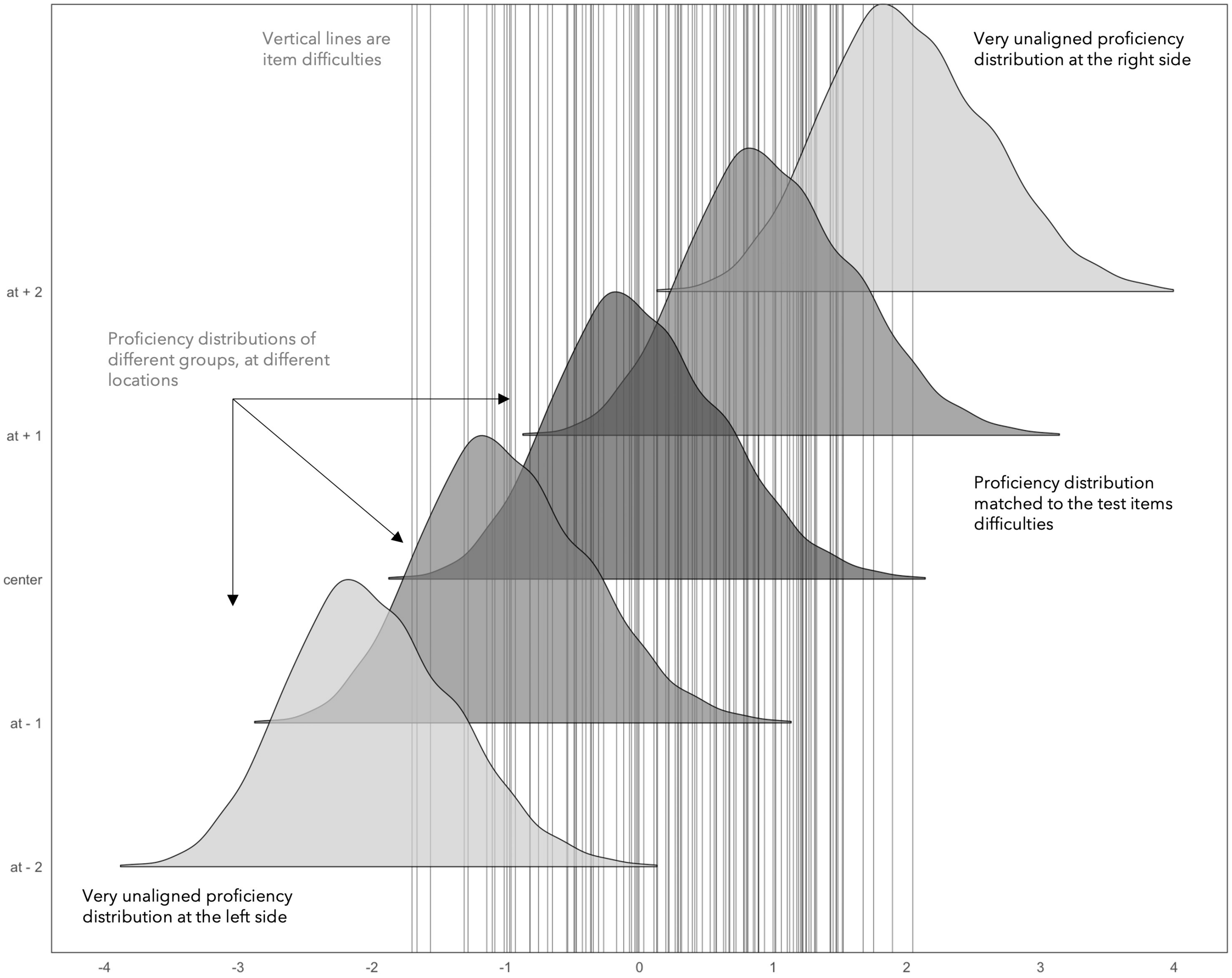
Test-Ability (miss)match

- In general terms we can imagine floor and ceiling effects. For a test to be informative for all participating countries the item difficulties must covered the full range of the different population abilities.

- In this sense, we think of this problem as an issue of test-ability (miss) match or test-ability alignment.

- **ILSA** is in the **most extreme scenario** to develop test that are comparable and informative for all target populations (high heterogeneity due to language, curriculum, and abilities). In contrast, Regional-ILSA may have lesser challenges due to lower language diversity, and lower curricular differences (hopefully).

- We assess what is the test-ability (miss) match in ERCE 2019, using Rasch Models.



Vertical lines are item difficulties

Very unaligned proficiency distribution at the right side

Proficiency distributions of different groups, at different locations

Proficiency distribution matched to the test items difficulties

Very unaligned proficiency distribution at the left side

ERCE 2013

Results

Test-Item difficulty alignment to participating countries abilities

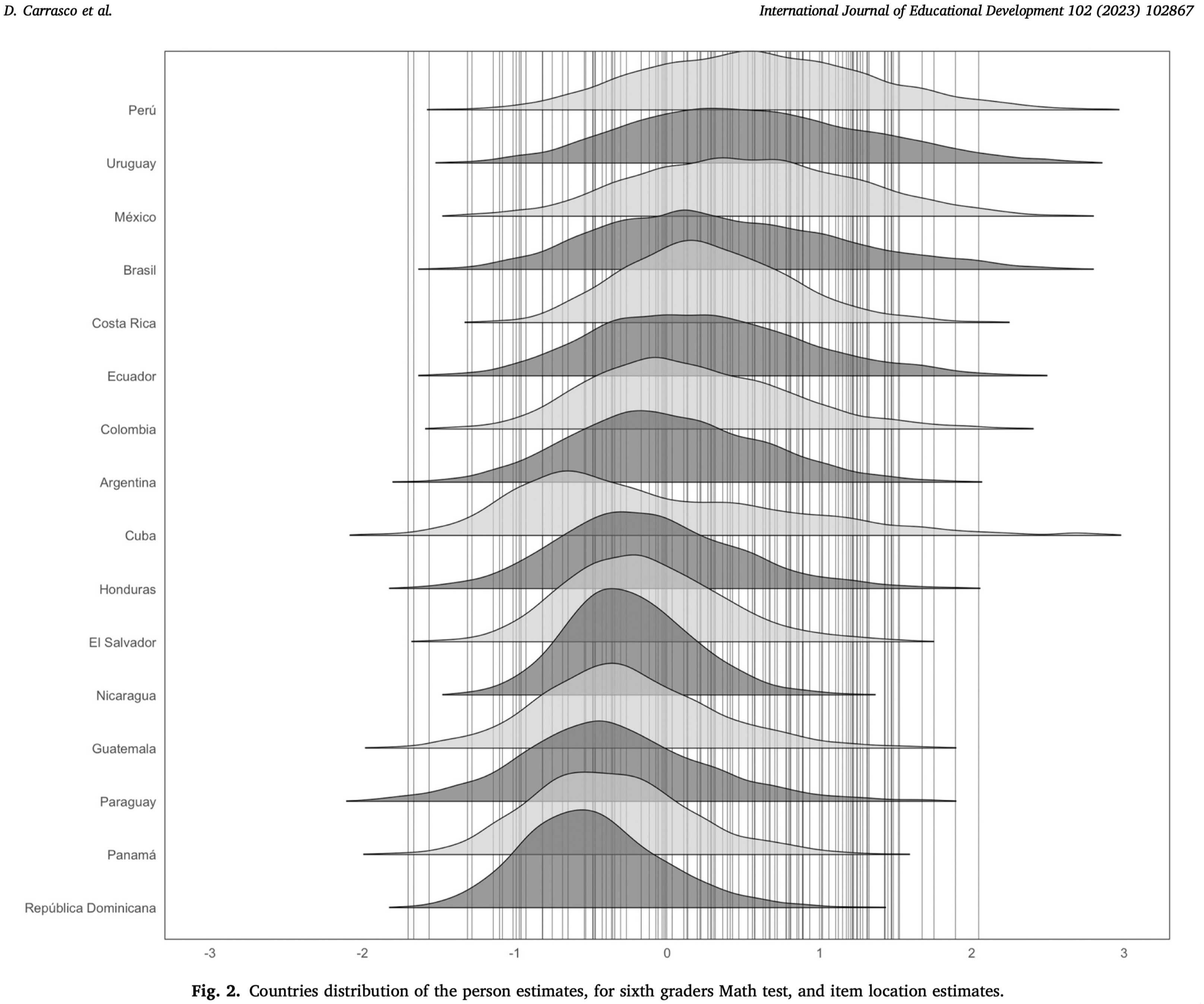
Item difficulty coverage

- ERCE 2013 test seems **well match** to the target participant abilities.

- There only fewer proportions of students outside the range of item difficulties. Similar results were obtained for 3rd grade (Math, and Language) and for 6th grade (Math, Language and Science).

- We believe these results are favored due to two main conditions: a **smaller heterogeneity** in language among participants, and smaller range of participating countries.

- Yet there are other features of ERCE **governance** and **design** we think is important to highlight: participation in test development, and curricular studies preceding test development.



D. Carrasco et al. *International Journal of Educational Development 102 (2023) 102867*

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

Rationale

Discussion

Uncertainty of scores at different scale locations

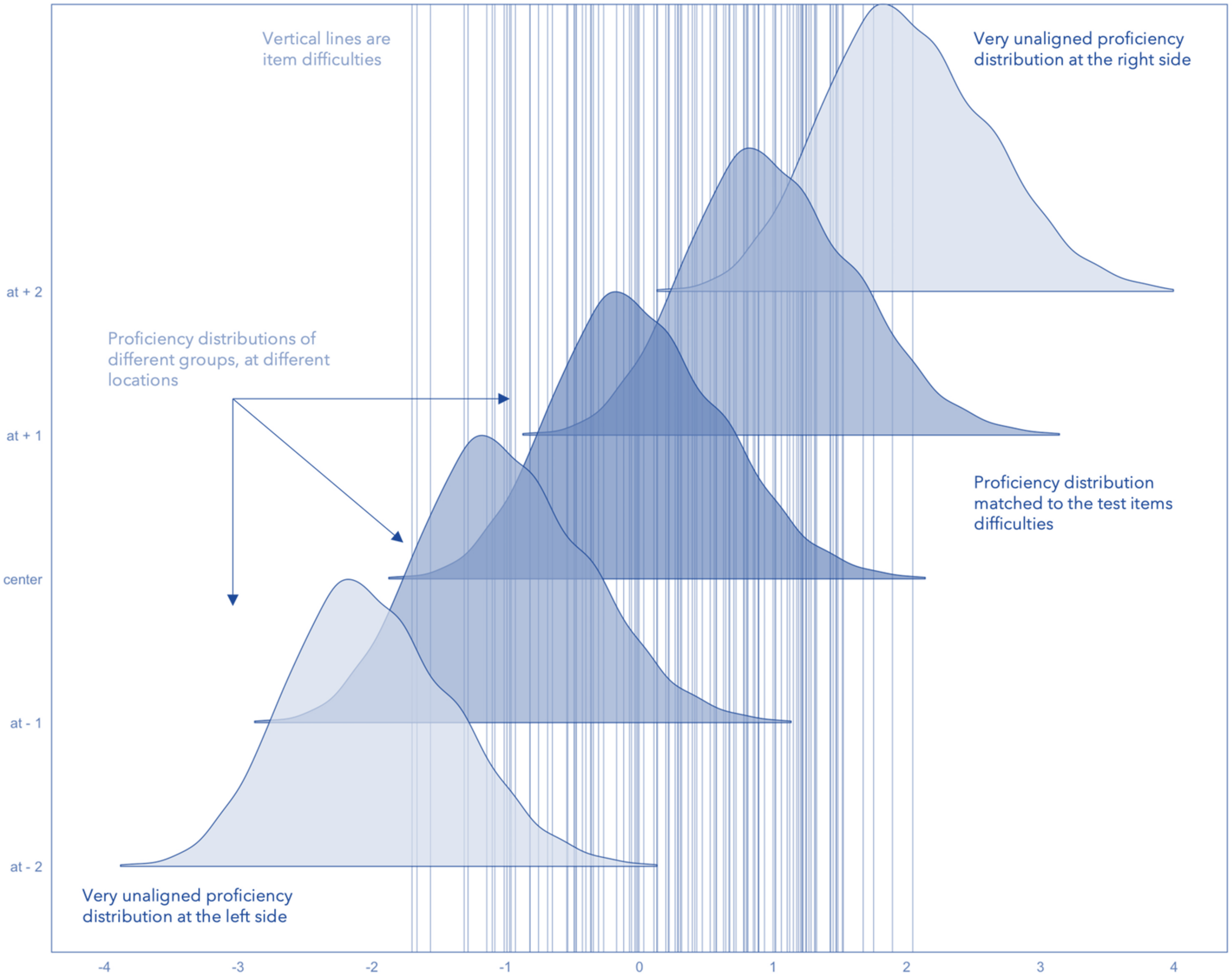
Discussion

- The higher the **heterogeneity** in **ability** between the countries we aim to assess, the larger must be the item difficulty coverage of the item bank we are using to assemble the tests.

- **Language diversity** among participating countries, increases the difficulty to create comparable item banks. ERCE 2013, only includes **two languages** (e.g., castellano-spanish, and Portuguese)

- **Curricular variability** poses challenges to assemble test with enough opportunities to learn among participating countries. 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

- **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.



Muchas gracias!

Carrasco, D., PhD
Centro de Medición MIDE UC,
Pontificia Universidad Católica de Chile
https://dacarras.github.io/

References

Carrasco, D., Rutkowski, D., & Rutkowski, L. (2023). The advantages of regional large-scale assessments: Evidence from the ERCE learning survey. International Journal of Educational Development, 102(May), 102867. https://doi.org/10.1016/j.ijedudev.2023.102867

Rutkowski, D., & Rutkowski, L. (2021). Running the wrong race? The case of pisa for development. Comparative Education Review, 65(1), 147–165. https://doi.org/10.1086/712409

Rutkowski, D., & Rutkowski, L. (2022). Designing Measurement for All Students in ILSAs. In International Handbook ofComparative Large-Scale Studies in Education (pp. 685–700). https://doi.org/10.1007/978-3-030-88178-8_27

Rutkowski, L., Liaw, Y. L., Svetina, D., & Rutkowski, D. (2022). Multistage Testing in Heterogeneous Populations: Some Design and Implementation Considerations. Applied Psychological Measurement, 46(6), 494–508. https://doi.org/10.1177/01466216221108123

Rutkowski, L., & Rutkowski, D. (2019). Methodological Challenges to Measuring Heterogeneous Populations Internationally. In The SAGE Handbook of Comparative Studies in Education (pp. 126–140). SAGE Publications Ltd. https://sk.sagepub.com/reference/sage-handbook-of-comparative-studies-in-education/i1259.xml

Rutkowski, L., Rutkowski, D., & Liaw, Y. L. (2019). The existence and impact of floor effects for low-performing PISA participants. Assessment in Education: Principles, Policy and Practice, 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 (pp. 749–767). https://doi.org/10.1007/978-3-030-88178-8_63

Tijmstra, J., Bolsinova, M., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2020). Sensitivity of the RMSD for Detecting Item-Level Misfit in Low-Performing Countries. Journal of Educational Measurement, 57(4), 566–583. https://doi.org/10.1111/jedm.12263

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