Technical Cooperation Group
9th meeting
22-24 November 2022

WG/GAML/6

Feasibility of using science scores from different educational assessments as proxy measures of reading literacy to measure and monitor SDG 4.1.1

Global Alliance to Monitor Learning (GAML)
23 November 2022

FEASIBILITY OF USING SCIENCE SCORES FROM DIFFERENT EDUCATIONAL ASSESSMENTS AS PROXY MEASURES OF READING LITERACY

TO MEASURE AND MONITOR SDG 4.1.1.

Daniel Miranda || Andres Sandoval-Hernandez
MIDE UC University of Bath

Introduction

This document evaluates several arguments related to the use of science scores as a proxy measure for reading literacy in the framework of measuring and monitoring SDG 4.1.1 at a global scale. The SDG indicator 4.1.1 measures the proportion of children and young people (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex. The arguments presented here are evaluated in relation to problems associated with the validity of these measures when scores from different subjects (e.g., literacy, science) are used interchangeably. Our evaluation is focused on three aspects: problems associated with differences in the conceptual framework on which the different tests are based; problems associated with the different interpretations that it is possible to make of the scores analysed, and problems associated with the relevant differences that are observed when these measures are correlated with student background factors such as gender.

Multiple efforts have been made to define standards that establish the quality of educational assessments. The main standards refer to the validity, reliability and fairness of the tests. The Standards for Educational and Psychological Testing defines validity as the "degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (EERA et al., 1999, 2014). This definition reminds us that one of the main aspects of validity has to do with what we intend to do with the results of the tests (i.e., the scores). In that sense, the validation process involves accumulating relevant evidence to provide a scientific basis for the proposed interpretation of the scores. In other words, "a clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided". Based on these definitions, a series of specific considerations regarding the theoretical framework and the possible uses of the scores from International Large-Scale Assessments (ILSA) to measure SDG 4.1.1 are derived. Arguably, the most relevant are the following:

- **Conceptual framework:** the conceptual definition of the construct(s) the test intends to assess. The construct or constructs that the test is intended to assess should be clearly described.

- **Intended interpretation:** test developers intended interpretation and use of test scores (e.g., disaggregation). A rationale should be presented for each intended interpretation of test scores for a given use.

- **Correlation with relevant factors:** the association of different test scores to a relevant background/sociodemographic factors (e.g., gender).

Conceptual frameworks

In PISA, for example, reading literacy is defined as "understanding, using, evaluating, reflecting on and engaging with texts to achieve one's goals, to develop one's knowledge and potential and to participate in society", while scientific literacy is "the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen” that “is willing to engage in reasoned discourse about science and technology which requires the competencies of explaining phenomena scientifically, evaluating and designing scientific inquiry and Interpreting data and evidence scientifically”. As can be observed, the general definitions offered in the PISA Assessment and Analytical Framework (OECD, 2019a), for each of the constructs present important differences in terms of what they intend to evaluate. While reading literacy focuses on comprehension processes, text format and situations or purposes; scientific literacy emphasises the contexts, types of knowledge and competencies.

Differences in the definitions provided in the assessment framework are evident. Reading literacy evaluations consider comprehension processes, text formats and situations. Comprehension processes refer to proficiency in locating information (e.g., accessing, retrieving and searching information in a text), understanding texts, and evaluating and reflecting on texts. Text formats can be static/dynamic, (non)continuous, or mixed texts (a combination of two or more formats). Finally, the framework considers the situations of use for which the texts were constructed (e.g, novels, personal letters or official texts). On the other hand, the definition provided for science literacy considers contexts, knowledge and competencies. Regarding the contexts, personal, national/local and historical issues are included. Furthermore, knowledge distinguishes between content knowledge (knowledge about the natural world and technological artefacts), procedural knowledge (knowledge about how ideas are produced), and epistemic knowledge (knowledge about the underlying rationale for the production of knowledge and the justification for its use).

Finally, the evaluation of scientific literacy emphasises the ability to explain phenomena scientifically, evaluate and design scientific searches, and interpret data scientifically (OECD, 2019a).

The differences between the conceptual frameworks of each test are evident. They have different purposes, different evaluative domains, and different skills are put into practice when facing the evaluation; they also define different competencies necessary to face their respective evaluation tasks and different degrees of complexity in the conceptualization.

Intended interpretations

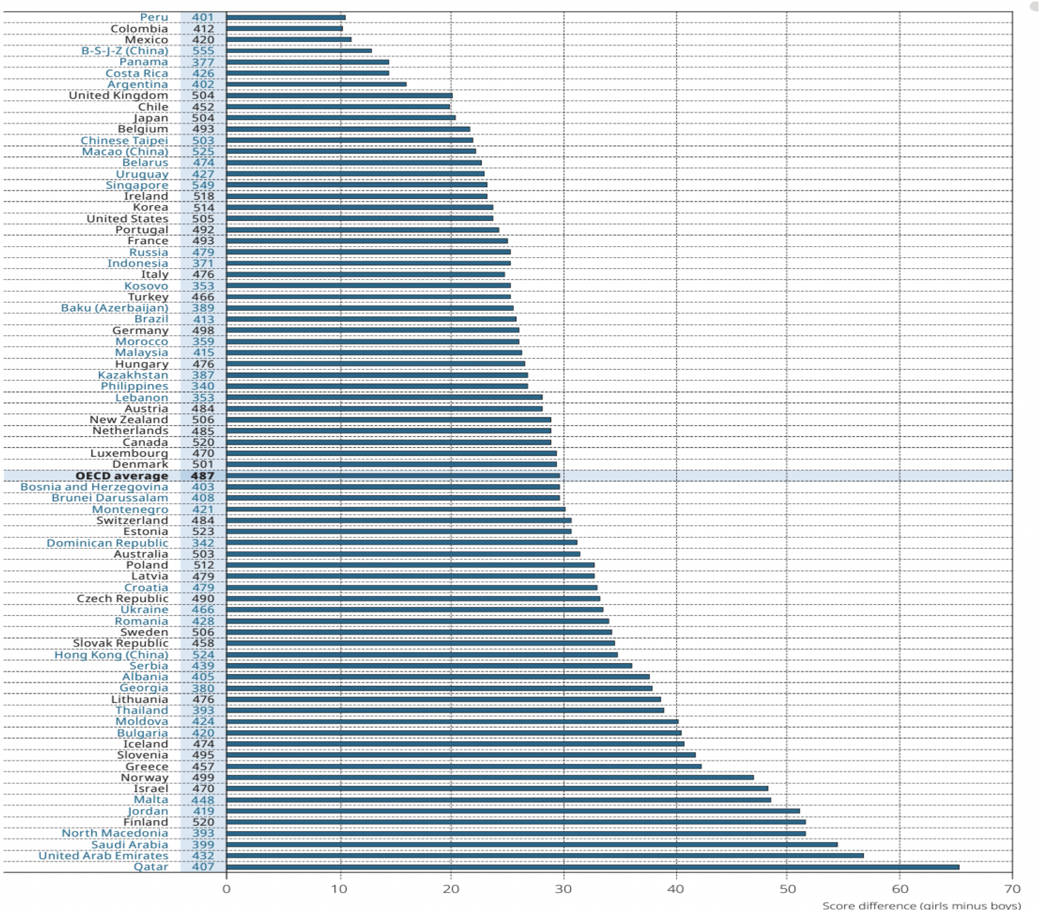
International large-scale assessments developed to assess reading, maths or science literacy are based on conceptual separate frameworks and, therefore, have different intended interpretations of the scores. In these assessments, scores are constructed based on the level of difficulty of the test and the level of ability of the students. Moreover, the estimation of the scores is normally disaggregated into subdomains that each test is designed to measure. In the case of reading literacy in PIRLS, for example, scores are also estimated for two subdomains based on *comprehension processes*: Retrieving or Straightforward Inferencing and Interpreting, Integrating and Evaluating (Mullis & Martin, 2019). On the other hand, in TIMSS, the subdomains measured in science are based on cognitive domains and are very different from those measured in reading: Knowing, applying and reasoning (Mullis & Martin, 2017).

Although to solve the questions included in a science test some reading literacy skills are put into practice, the questions included in international large-scale assessments are designed to avoid this dependency. For example, when describing the science test, one of the most recent PISA reports mentions that “to address these concerns, stimulus material and questions use language that is as clear, simple, brief and syntactically simple as possible while still conveying the appropriate meaning. The number of concepts introduced per paragraph is limited. Questions within the domain of science that specifically assess reading or mathematical literacy are avoided” (OECD, 2019b, p. 113). Therefore, the assessment is primarily and specifically oriented towards assessing the conceptual framework of science literacy rather than other forms of achievement, such as reading literacy.

Correlation with other relevant factors

An additional way of testing whether two tests or instruments measure the same construct is to observe how their scores correlate to other relevant factors. For example, when analysing the results in previous versions of PISA, it can be observed that the achievement scores in the different domains (i.e., mathematics, science, reading) show different patterns in their associations with sociodemographic factors (e.g., socioeconomic status or student's sex). Figure 1 shows the gender differences in reading performance.

Figure 1. Gender gap in reading performance in PISA 2018.



Notes: The mean score in reading is shown next to the country name.

All differences are statistically significant.

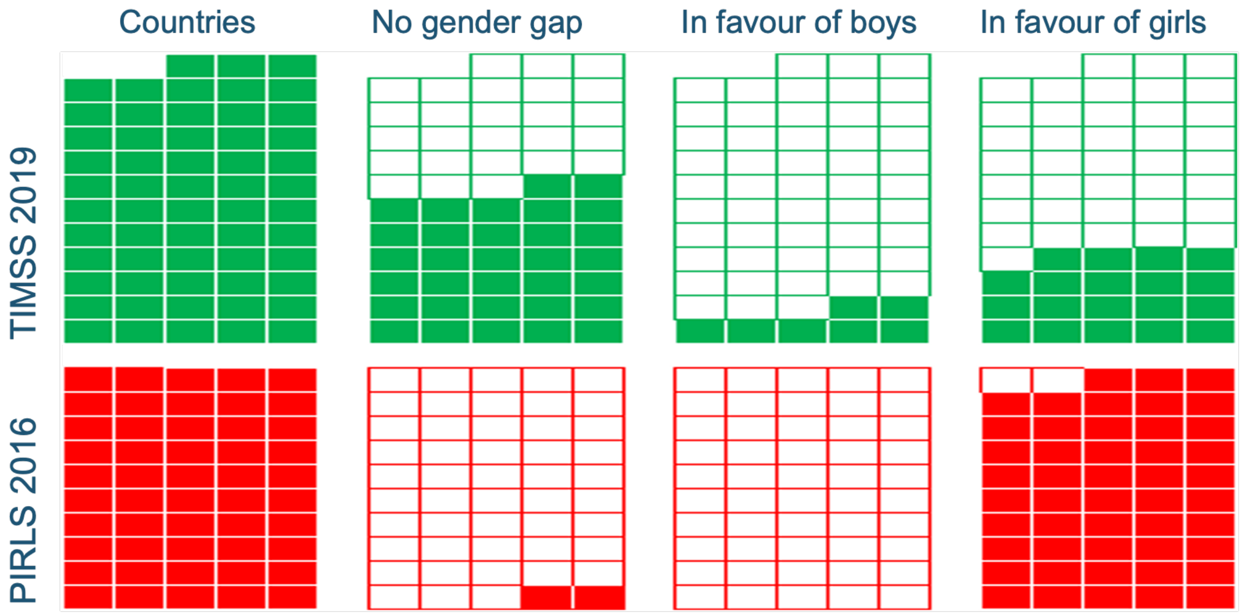
Countries are ranked in ascending order of the score point difference related to gender (girls minus boys).

Source: (OECD, 2019b, Tables I.B1.4 and II.B1.7.1)

As can be seen in Figure 1, the result always favours girls over boys, on average, by 29 points. Additionally, it can be observed that there are important differences between countries in these differences. For example, in Latin American countries the gaps are smaller, while in Arab countries and some developed countries the differences are much more significant. In contrast, for science performance, the differences between boys and girls are much smaller. On average across OECD countries in 2018, girls outperformed boys in science by two score points; and in around half of the participating countries, the performance difference between boys and girls was not statistically significant. In only 6 countries boys' performance in science was significantly higher than that of girls; but the opposite was observed in 34 countries and economies (OECD, 2019b).

An additional example can be made with data from IEA studies. Figure 2 shows the waffle chart, where each of the shaded squares corresponds to a country where gender gaps were found. The Figure also shows the number of countries where the gap is in favour of boys or girls, respectively.

Figure 2. Gender gap in reading performance in TIMSS 2019 and PIRLS 2016.



Notes: The graphs in the first column represent the number of countries that participated in each study. The second column shows the countries where the differences in average scores between boys and girls were not statistically significant. Columns 3 and 4 show the countries where the difference was statistically significant in favour of boys or girls, respectively.

Source: Own calculations based on TIMSS (Mullis et al., 2020) and PIRLS (Mullis et al., 2017).

As can be observed in Figure 2, in TIMSS and PIRLS, there are also important differences in the patterns observed in the associations between science and reading test scores and gender. While in TIMSS (science) significant differences were observed in less than half of the participating countries, in PIRLS (reading) these differences were observed in almost all countries (48 out 0f 50). Furthermore, in reading, all the differences were in favour of girls, while in science the differences favoured boys in seven countries.

These differences in the association patterns between reading and gender and science and gender have implications regarding the use of these scores as interchangeable. According to the Standards for Educational and Psychological Testing (AERA et al., 2014), where there is relevant evidence indicating that test scores may differ in meaning for relevant subgroups, its implications for the validity of the use of scores should be examined. Not considering these differences implies ignoring the potential consequences for fairness in using the tests. For this, it is necessary to delve into the under-representation of the construct or the identification of irrelevant variance of the construct. Additional analysis may consider the test content, internal structure, the relationship of scores with other variables, or an analysis of the response process.

Additionally, is well known that there is a correlation between achievement test scores. Mathematics achievement tends to show a high correlation with science achievement, as well as science scores tend to establish high correlations with reading literacy. The question that rise about this pattern is how can the shared variance between different types of literacy be interpreted? Can the fact that one test shares a proportion of variance with other be interpreted as those scores measuring the same construct?

These correlations are expected since ‘good’ students tend to be higher achievers across subjects. Students that achieve higher scores in mathematics tests tend to achieve higher scores in other subjects' tests. Nevertheless, the association between test scores does not mean that different test measure the same construct. On the contrary, “relationships between test scores and other measures intended to assess the same or similar constructs provide convergent evidence, whereas relationships between measures purportedly of different constructs provide discriminant evidence” (AERA et al., 2014).

Final thoughts

Based on the arguments presented here, the following considerations regarding the possible use of the scores from ILSAs to measure SDG 4.1.1 are derived.

- Test developers should clearly state how test scores are intended to be interpreted and consequently used.

- If validity for some common or likely interpretation for a given use has not been evaluated, or if such an interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.

- If a test score is interpreted in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence, if necessary.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS International Study Center and IEA.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). *PIRLS 2021 Assessment Frameworks*. TIMSS & PIRLS International Study Center and IEA.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2016/international-results/

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-results/

OECD. (2019a). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing.

OECD. (2019b). *PISA 2018 Results: Where All Students Can Succeed: Vol. Volume II*. OECD Publishing. https://doi.org/10.1787/b5fd1b8f-en