The feasibility of using science scores from different educational assessments as proxy measures of literacy

to measure SDG 4.1.1

Andrés Sandoval-Hernández[1]
Daniel Miranda[2]

[1]University of Bath, [2]Centro de Medición MIDE UC

unesco Institute for Statistics

GLOBAL ALLIANCE TO MONITOR LEARNING

November 23rd
2022

UNIVERSITY OF BATH

Centro UC Medición - MIDE

Contents

Contents

Proxy measures to monitor SDG 4.1.1

Introduction

Sustainable Developmental Goal 4.1.1

SDG 4.1.1

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

Starting points

- Can we use science scores to proxy reading scores when measuring SDG 4.1.1?
- Multiple efforts have been made to define standards that establish the quality of educational assessments → validity
- Validity: “the degree to which evidence and theory support the interpretations for proposed uses of test scores”
- The validation process involves accumulating relevant evidence to provide a scientific basis for the proposed interpretation of the scores

Starting points

- We propose that at least three points should be considered when answering this question:

  - **Conceptual framework:** the conceptual definition of the construct(s) the test intends to assess.

  - **Intended interpretation:** test developers should clearly set forth how test scores are intended to be interpreted and used (e.g., disaggregation).

  - **Correlation with relevant factors:** it is incumbent on the user to justify any new/unintended interpretations of test scores, provide a rationale, and collect new evidence, if necessary.

Proxy measures to monitor SDG 4.1.1

Conceptual frameworks

Differences in definitions

- The case of PISA as an example:

- **Reading**: “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”

- **Science:** “the ability to engage with science-related issues, as a reflective citizen […] which requires the competencies of: explaining phenomena scientifically, evaluating and designing scientific inquiry and Interpreting data and evidence scientifically”

Differences in what is considered in the evaluation

- **Reading**: comprehension processes, text format and situations or purposes.

- **Science:** contexts, types of knowledge and competencies.

- Differences in the contextual frameworks are evident (e.g., purpose, evaluative domains, skills, competencies, etc.)

Proxy measures to monitor SDG 4.1.1

Intended interpretation

Differences in the composition of the main construct

- The case of IEA (TIMSS/PIRLS) as an example

- **Reading:** two comprehension processes: Retrieving/Straightforward Inferencing and Interpreting/Integrating/Evaluating

- **Science:** three cognitive domains: knowing, applying and reasoning

Differences in the composition of the main construct

- Although solving science items requires reading comprehension, science tests are designed to avoid this dependency

- The latest PISA report mentions:

  - *“to address these concerns, stimulus material and questions use language that is as clear, simple, brief and syntactically simple as possible”*

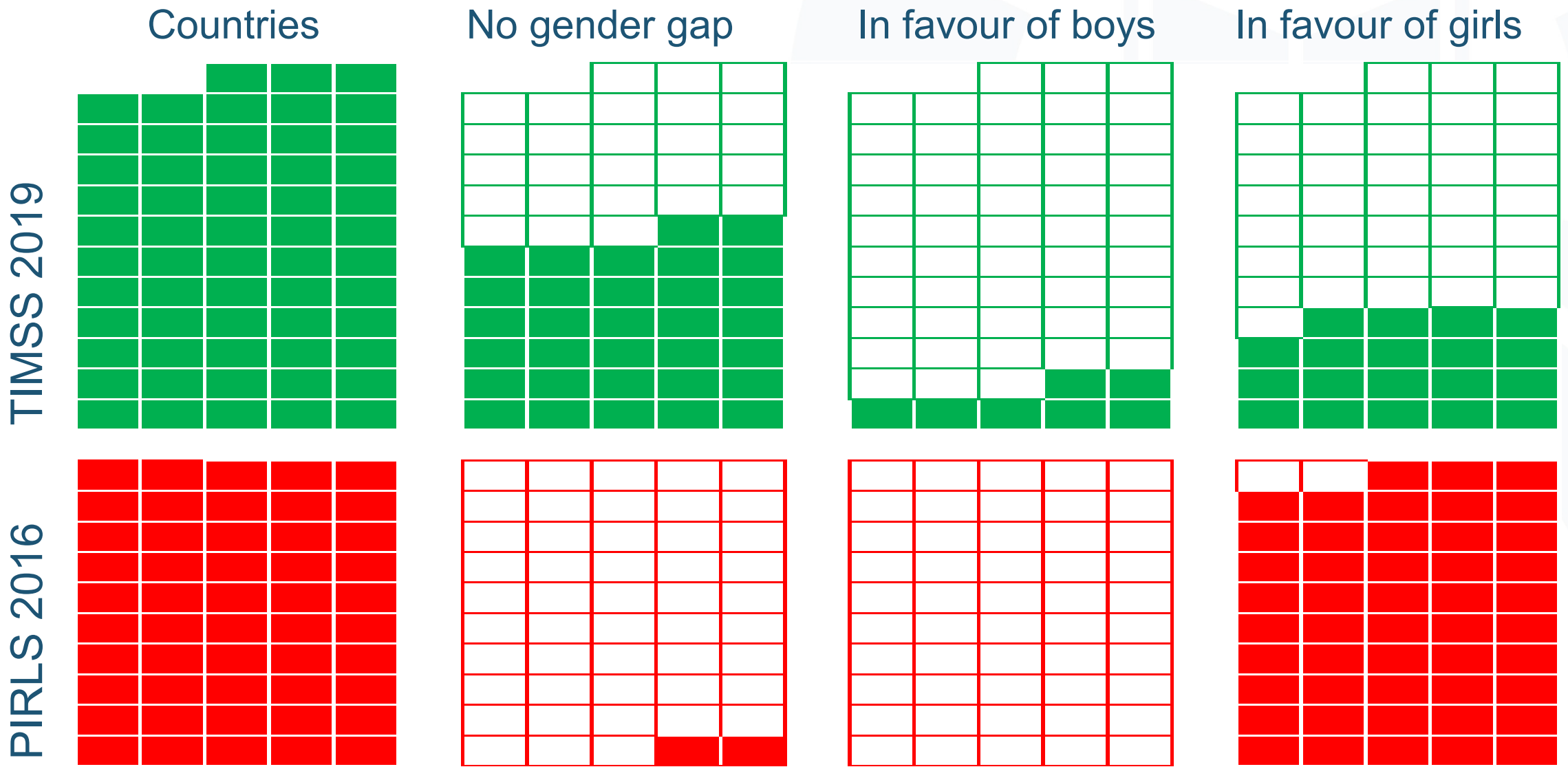
  - *“The number of concepts introduced per paragraph is limited”*

  - *“Questions within the domain of science that specifically assess reading, or mathematical literacy are avoided”*

Proxy measures to monitor SDG 4.1.1

Correlations with relevant factors

Gender gaps


Countries
No gender gap
In favour of boys
In favour of girls
TIMSS 2019
PIRLS 2016

Correlations/associations with sex

- An additional way of testing whether two tests measure the same is to observe how they correlate to relevant factors.

- According to the Standards for Educational and Psychological Testing, 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, or biased policy recommendations.

Proxy measures to monitor SDG 4.1.1

Conclusions

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.

Proxy measures to monitor SDG 4.1.1

Questions & Comments?

Thank you!

Andrés Sandoval-Hernández

University of Bath

Daniel Miranda

Centro de Medición MIDE UC,

Pontificia Universidad Católica