Floor effects and the comparability of developing country student test scores

Martin Gustafsson and Bilal Barakat

February 2020


**Floor effects** can be understood as non-differentiation among worse performing students, because the test is set at too difficult a level, relative to the competencies of students. If too many students are ‘under the floor’, meaning they get nothing right in the test, and obtain a score of zero, it becomes difficult to compare results, for instance across countries in an international programme, or within countries over time.

**Things we already knew:**
- Floor effects create comparability problems, and result in there being little information on the most marginalised students. There is little information on what they can and cannot do.

**Things we needed to understand better:**
- How to determine the extent of floor effects in a testing system, given realities such as the use of multiple choice (MC) questions, which allow for random guessing.
- The effect of floor effects on comparability within international testing programmes commonly used in global monitoring efforts.

**Burton (2001). Quantifying the effects of chance in multiple choice and true/false tests**

An analysis of theoretically derived data.

**Rutkowski et al (2019). The existence and impact of floor effects for low-performing PISA participants**

An analysis of actual and simulated PISA data. (No specific analysis of random guessing effects.)

Many MC tests would display random guessing effects which are too large for the tests to fulfil their intended purpose.

**OUR ANALYSIS**

Though regional programmes such as LLECE and SACMEQ were designed to be more suitable in developing country contexts, they display substantially higher floor effects than TIMSS, which has in recent years introduced easier testing for developing countries. This problem is largely due to LLECE and SACMEQ being based almost exclusively on MC questions, while half of all questions in TIMSS are non-MC ‘constructed response’ questions. After taking into account uncertainties relating to floor effects, across-country rankings in LLECE and SACMEQ hardly change, but monitoring progress over time would be seriously compromised.

The best way of tackling the problem, and ensuring that better information on more marginalised students is generated in LLECE and SACMEQ, may be to introduce more easy MC questions. More constructed response questions raise the costs and complexities of testing, meaning this route may not be feasible.

**An adaptation of Burton’s work, for instance to distinguish between incorrect MC response and missing MC response. An application of the method to item-level TIMSS, LLECE and SACMEQ data (all from the primary level) to gauge the extent of floor effects and its impact on comparability.**

The graph on the left captures Burton’s theory: A student who obtains, say, 40 out of 100 in a purely constructed response (CR) test, would obtain about 50 in an MC test due to random guessing, which also comes with probability distributions.

The graph on the right converts student-level CR scores (vertical axis) to MC scores, with margins of error for the latter. Four-choice MC questions assumed.

**In SACMEQ 2007, Grade 6, virtually all students obtain a classical score of at least one correct. However, after a conversion of student scores to a result which removes the benefits of random guessing (while differentiating between missing MC responses and incorrect MC responses), one sees large proportions of students scoring zero: up to 37% in Zambia’s mathematics. A similar analysis of LLECE Grade 6, from 2013, reveals a very similar problem.**

Ignoring floor effects results in an over-estimation of students reaching minimum standards. For instance, in Dominican Republic, 12% of students officially do not meet a minimum level of proficiency. This becomes 37% after the benefits presented by random guessing are removed.

The 12% and 37% statistics would display similar confidence intervals.