Floor effects and the comparability of developing country student test scores

May 2019

Martin Gustafsson
Research on Socio-Economic Policy (ReSEP)
Department of Economics
University of Stellenbosch

Abstract

To an increasing extent international assessment results inform education policy debates, yet little is known about the floor effects in these assessments. To what extent do they fail to differentiate between the most disadvantaged, and what are the implications of this, for instance in terms of the comparability of national statistics across space and time? Microdata from TIMSS, SACMEQ and LLECE are analysed to answer this question, with reference to primary schools. In TIMSS, floor effects have been greatly reduced through the introduction, in 2015, of TIMSS Numeracy, which includes a greater number of easier items relative to regular TIMSS. SACMEQ and LLECE, despite being specifically designed for developing countries, often display large floor effects. This results in a situation where many students scoring zero, after adjustments for random guessing, are classified as having passed proficiency thresholds. Though these floor effects do not substantially alter the rankings of countries, they are large enough to undermine proper monitoring of progress over time. They can also undermine public trust in the programmes, and they leave information gaps in relation to those students requiring most support. Designers of assessment programmes need to limit floor effects though the presence of more easy multiple choice items and more constructed response items. The former solution is the easiest to implement.

Contents

1 Introduction

There are many compelling reasons today for every country to participate in international assessments or run national assessments, one reason being that countries have agreed to report on progress in the proficiency of school students in line with the UN's Sustainable Development Goals (SDGs). Decisions around which international assessment programmes to participate in, what improvements or adaptations in these programmes countries should argue for, and how to design a national assessment programme, are complex, in part because these systems seldom fulfil just one purpose. The current paper focusses on just one aspect of assessment systems, namely 'floor effects'. The paper thus feeds into the wider body of knowledge required to make sound decisions around participation in and the design of assessment programmes.

Floor effects arise when the difficulty of an assessment is set at too high a level to produce meaningful information about the performance of academically weaker students. The result is that a large proportion of students at the low end of the performance spectrum all appear to be similarly weak, possibly without the ability to respond correctly to anything in the assessment. Floor effects produce at least two serios problems. Firstly, they make it difficult to determine what interventions are needed for the weakest students, because one knows so little about what they can and cannot do. Secondly, floor effects can distort aggregate statistics, such as those needed to monitor the SDGs. More broadly, accepting serious floor effects in assessment systems can be seen as conceding to a widespread but undesirable trend whereby the curriculum and teaching 'aim too high', because they are designed with just better performing students in mind. Floor effects are one manifestation of what Pritchett and Beatty (2012) refer to as the problem of over-ambitious curricula in developing countries.

Section 2 discusses the role of assessments broadly. What are the aspects of an assessment system a decisionmaker needs to take into account? What are the links between these and the purposes of standardised student assessments? How can assessment systems play a role in reducing inequalities and advancing inclusion? The purpose that receives most emphasis in the current paper is reporting in terms of the SDGs. However, it is also emphasised that this is but one of many purposes.

Section 3 discusses attempts to link different assessment programmes to produce comparable national educational quality statistics, for instance for the purposes of the SDGs. The benefits of these attempts for global organisations such as UNESCO, but also country-specific stakeholders, are discussed, as are risks associated with insufficient comparability. How floor effects might worsen comparability is discussed.

Section 4 explores the limited literature dealing with floor effects. Above all, an approach to adjusting classical test scores downward in order to take into account random guessing in tests with multiple choice questions is examined.

Sections 5 and 6 present the empirical work of the paper. The focus is largely on how to gauge the seriousness of floor effects in specific countries participating in a few international programmes. Seriousness is gauged partly in terms of the reliability of comparisons across countries, though comparability over time is also discussed. The analysis includes measuring random guessing effects and examining the impact of having only multiple choice questions. Section 5 uses TIMSS Grade 4 mathematics data from 2015 to establish useful methods. Section 6 then applies these methods to two programmes focussing on developing countries: Africa's SACMEQ[1] and Latin America's LLECE[2]. The focus is exclusively on the primary level. In part, the analysis is aimed at 'demystifying' item response theory (IRT) scores, which are now

[1] Southern and Eastern Africa Consortium for Monitoring Educational Quality.
[2] Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (Latin American Laboratory for Assessment of the Quality of Education).

widely used in international and national assessments, yet insufficiently understood by those working outside the field of psychometrics. There is some discussion of the extent to which children are not covered by the available test data, for instance because they are not attending school.

Section 7 offers a conclusion directed, in particular, at those involved in decision-making relating to assessments in developing countries.

2 Assessment purpose and design

Of the three kinds of assessment defined by Clarke (2012), only one receives attention in this paper: 'large-scale, system-level assessments'. The other two kinds, 'classroom assessments' and 'examinations', are not covered here for the simple reason that they cannot be used to report against the SDGs (though some might argue that examinations can – see UIS [2018: 24]). Reporting against the SDG indicators on learning proficiency is by no means the only purpose of large-scale system-level assessments, referred to as 'systemic assessments' henceforth. If that were the only purpose, it might be difficult to justify their costs. As explained below, there are several other purposes.

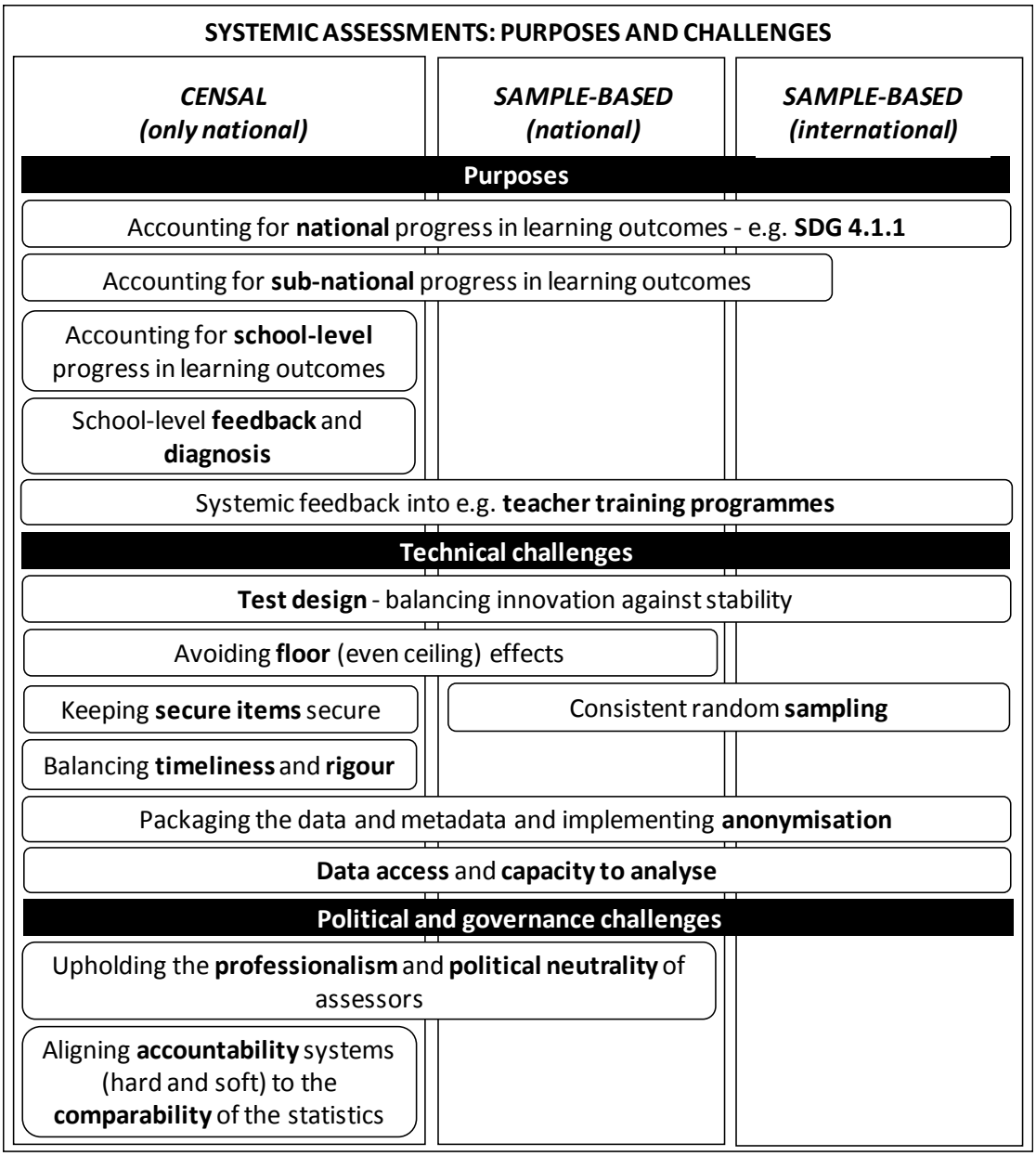
The diagram below captures the purposes and challenges of systemic assessments. Floor effects are but one of several challenges. There are many guides for policymakers taking decisions about what assessment systems to use – see for instance Clarke (2012) and the 2018 *World Development Report*. What is different in the discussion provided here, is that a clear distinction is made between censal (or universal) and sample-based systemic assessments. The implications of this distinction have arguably not received enough attention in the guides, yet if one talks to policymakers dealing with the design of national assessments, this is a critical issue.

Whether a systemic assessment is national or international, censal or sample-based, it has the potential to produce national statistics for an indicator such as SDG 4.1.1, which reads as follows:

> 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 emphasis in the diagram is on monitoring *progress*. One does not just want to know that one's indicator values are low, one wants to know whether they are improving over time, and whether the trend can be trusted. National systems are generally designed to monitor trends in sub-national units such as provinces. International assessment systems, on the other hand, are often poorly suited for this as sample sizes tend to be too small. Obviously, only censal systems covering all schools have the potential to serve as an accountability tool for individual schools, and as a source for preparing school-specific diagnoses of learning problems intended for school actors. However, sample-based systems can also generate useful diagnoses and feedback, such as specific information to designers of teacher training programmes on what teaching competencies require more attention.

Figure 1: Purposes and challenges of systemic assessments



Assessment design manuals pay considerable attention to appropriate test design. Perhaps the most useful 'manuals' are the test design reports of large and well-resourced assessment programmes such as TIMSS and PIRLS. To illustrate, *Methods and procedures in PIRLS 2016* describes the work required to update assessment frameworks, determine the right mix of multiple choice and constructed response items, develop scoring guides for the constructed response items, translate items and materials into the required languages, field test new items, combine different sets of items in different but equivalent tests in line with a matrix sampling approach, and to manage experts with varying opinions during a process lasting three years (Martin *et al*, 2017). Doing all this properly is clearly a lengthy and costly project. Experts tend to underline the need for consistency over time in test design procedures to allow for the comparability of results. However, this can be difficult for developing countries with limited capacity. In fact, countries seldom succeed in implementing the full set of best practices at the inception of a programme. Instead, innovations and improvements tend to be phased in over several waves of the assessment. There is thus a trade-off between maintaining a consistent test design approach and bringing about necessary innovation.

If newly designed assessment systems are geared towards the levels of proficiency seen in the student population, then floor effects can be avoided. A typical manifestation of floor effects is when test items are so difficult that large numbers of students fail to get any questions right, other than through random guessing in the case of multiple-response items. This should not happen. In psychometric terms, items should have sufficient discriminatory power (Izard, 2005) and they should display sufficient inclusiveness (UIS, 2017a: 7). It is highly advisable in a test written by, say, Grade 6 students to include some items which would typically be associated with grades below Grade 6. In particular in developing countries, skills specified in the curriculum for a specific grade tend to be well above what students in that grade actually get to learn (Pritchett and Beatty, 2012). A test focussing only on competencies specified in the Grade 6 curriculum is thus likely to be too difficult for a large number of Grade 6, or even Grade 7, students, leading to floor effects. As will be seen in section 5, ceiling effects, where a large number of students get many items right, also exist, though this is less common. How international testing systems such as TIMSS have dealt with floor effects in developing countries by producing parallel tests with more easy items is discussed in section 6.

Where systemic assessments are sample-based, random sampling approaches must be consistent so that, for instance, middle class students are not over-represented in a specific year, which is likely to bias upwardly the results. If background questionnaires are administered to students, it becomes possible to check the randomness and consistency of the sampling in powerful ways. Of course, background characteristics need not be static. If the middle class has in fact grown, this should be reflected in the sample.

A critical matter on which too little guidance exists is the following: In the case of a censal national assessment, how can secure items be kept secure? The only way to make test results comparable over time is to repeat test items, and if items are repeated, they need to be kept secure, or secret, at least until after the first repetition. This is not too difficult in a sample-based assessment, where there are a limited number of schools and it is possible to employ enough test administrators to prevent leakages. Two problems arise in the case of censal assessments. Firstly, the large number of schools and people involved increases the risk of leaks. Secondly, censal systems are inevitably seen by schools as an evaluation of each school's performance, as opposed to a distant sample-based research project, meaning the incentive to keep tests for practice purposes among school principals and teachers is high. Yet several countries appear to run censal assessments with secure items shared across years successfully: Chile, Brazil and Australia, to name a few (World Bank, 2017: 92-93). Are these countries really successful in keeping secure items secure? If they are, how do they do it? These questions are vital for policymakers, who are often interested in assessments covering all schools, in part so that accountability to parents and the administration can be strengthened.

Australia's National Assessment Programme (NAP), which is particularly well documented, describes what is probably the only viable solution. NAP tests all students in specific grades across the country using exactly the same test within a grade and subject, in a specific year. Each year completely new tests are developed, with no items common across years in the universal testing. In parallel, testing of a very small 'equating sample' occurs each year, using some test items from the universal tests, but also completely secure items repeated across years in the equating sample. By comparing the results of the same students in the universal test and the equating sample test, and adjusting universal test results, it is possible to produce, for each school, average scores which are sufficiently comparable over time. The data collected through the equating sample serves as the common thread over time. It does not matter if tests from the universal testing become public after everyone has written the test in a year. The equating sample requires only around 1,000 students each year, across around 40 schools, in the case of each grade and subject. Testing of the equating sample occurs before the universal testing each year (Australian Curriculum, Assessment and Reporting Authority, 2017: 18-19).

The magnitude of censal assessments creates serious logistical challenges. Large volumes of data must be collected, and analysts need enough time to produce statistically valid final results, results which must inevitably take the form of item response theory (IRT) scores as adjustments of the raw data are needed (more on IRT in section 5). At the same time, in censal assessments there is much pressure for results to be fed back to schools quickly before the statistics become irrelevant for comparison and remediation purposes. The assessment authority would need to make difficult decisions about timing. Rigour in relation to the data cannot be unduly compromised, yet schools should not wait for too long. Australia's NAP, a highly developed system, achieves a turnaround period of four months between the writing of the universal tests, which occurs online for some schools and on paper for others, and the release of results to schools. Clearly, the complexity of running censal assessments *which are comparable over time* should not be under-emphasised. The case for relying solely on a sample-based programme for gauging progress over time is a strong one for developing countries with limited technical capacity.

International programmes such as TIMSS provide researchers with access to the assessment and background questionnaire microdata. This is also the case in many of the better run national assessments. The advantage with this access is that the more researchers there are working on the data, the more knowledge is generated, and hence the better informed the national policy debates. The potential for knowledge generation of most assessment datasets is immense, way beyond what can be generated by a few researchers working in the national authority. In Chile, to take an example, microdata from the national censal assessment programme, SIMCE, are available to researchers. However, making data accessible comes with its own complexities. Proper metadata (technical documentation) must be prepared, and the identifiers of individual students and teachers (and perhaps even schools) may have to be anonymised to in line with national policies on data privacy.

At the political and governance level, a key challenge is to protect the independence of the national assessment authority. Assessment results are politically sensitive as they can be used to signal the success or failure of governments. The authority should not only be free from political interference, it should be *seen* to be acting independently.

Finally, in the case of censal assessments, the temptation should be avoided of 'loading' the programme with greater accountability significance than what is justified, given the technical rigour of the programme and the comparability of results. Any assessment programme, in particular a new one, is likely to be limited in terms of its ability to truly tell whether schools are improving or deteriorating over time. If comparability is weak, or even seen to be weak, using results from the assessment to make hard determinations around which schools are exemplary, or which schools need remediation, may not be justified, or could be seen as unfair.

3 Attempts to link assessments

The current paper focusses largely on floor effects and what they mean for comparability across countries and, to some extent, over time *within* an international testing programme. However, an additional concern should be the extent to which floor effects compromise comparability *across* testing programmes where various programmes have been linked. Of particular relevance here is the recent work of Altinok, who has linked programmes to produce 'harmonised' global datasets for both the UIS and the World Bank (Altinok, 2017; Altinok *et al*, 2018).

But before comparability across testing programmes within harmonised datasets is discussed, it is worth asking the question of whether comparability is what it should be *within* programmes. Work such as that of Altinok implicitly makes the assumption that comparability is good within programmes. If this assumption is not true, that creates additional problems for harmonised datasets, as any within-programme discrepancies would be replicated within such datasets.

In fact, comparability within programmes, while apparently good, is often not as good as analysts from within these programmes claim. There is a limited literature on within-programme comparability programmes. Prominent here is an article by Jerrim (2013), who argues irregular sampling procedures in England's PISA testing produced serious performance declines which may not be real. Carnoy *et al* (2015) argue that sampling inconsistencies in the case of Brazil meant that annual improvements in PISA were to some extent exaggerated. Clearly, apart from floor effects, issues such as sampling can compromise comparability within a programme.

Adjusting scores from various international assessment programmes to create harmonised country-level datasets has been pursued by a small handful of analysts. This kind of work seems to have begun with Hanushek and Woessman's (2009) combining of data from PISA, TIMSS and a few programmes other than TIMSS run by the IEA[3]. Hanushek and Woessman's intention was to produce a dataset which would explain the economic growth of countries in a more robust fashion. Their finding of a very clear relationship between educational quality and economic growth is part of the reason why the education SDGs pay such close attention to learning outcomes.

Altinok's work, which draws heavily from the earlier Hanushek and Woessman work, aims to provide historical statistics and linking methods which assist in the monitoring of SDG 4.1.1. Altinok focusses both on the means and proportion proficient of each country. He has drawn from 15 international testing programmes and produced statistics for around 150 countries for various points in the 1965 to 2015 period. He argues that these statistics are sufficiently comparable for SDG monitoring purposes.

Given the political sensitivity of across-country comparisons, and especially trends over time, trends which can influence elections, one can expect there to be a growing body of research refuting and defending the comparability of Altinok's statistics, and other similar sets of harmonised statistics which are likely to emerge in future. There is already some work questioning the feasibility of linking different programmes, such as that of Hastedt and Desa (2015), but this type of work is still in its infancy. Proper critiquing will obviously require access to the methodological details and resultant statistics of work that has been undertaken. Though the statistics arising out of Altinok (2017) are available[4], when the current report was produced the harmonised dataset described in Altinok *et al* (2018) was not yet publicly available. However, the 2018 report itself contains a table with means per country, where each mean could span many years.

Likely concerns about the linking methodologies employed by Altinok would fall into two categories. Firstly, there could be concerns that the 'doubloon' countries used to link different assessment programmes display rather different samples across the two programmes. To illustrate, Altinok *et al* (2018: 37) use Colombia as a doubloon country to link LLECE 1997 Grade 6 mathematics to TIMSS 1995 Grade 8 mathematics. Linking two different grades in this manner is not necessarily problematic, if the samples are similar in terms of factors such as the percentage of out-school-children and grade repetition, and if different countries display similar grade-on-grade improvements in learning outcomes. Clearly, such assumptions will tend to be violated, though the critical question is to what extent. And if the aim is to produce comparable proportion proficient statistics, then having two different grades is likely to be problematic as this statistic can be expected to change systematically with grade, even within the same country (UIS, 2018: 17). Secondly, a doubloon country may have a sample which is somehow distorted relative to other countries in the same programme. Had England been used

[3] International Association for the Evaluation of Educational Achievement.
[4] See the 'UIS Learning Outcomes Anchored Database' at http://tcg.uis.unesco.org/data-resources (accessed May 2019).

as a doubloon country when England had an irregular PISA sample, that would weaken the link between another programme and PISA.

A further and fundamental area of concern can be raised. Clearly, the original data must truly represent each country. This is not the case in Altinok's harmonised dataset for the world's two most populous countries. For both China and India, PISA data for just a few regions, which are unlikely to be representative of the country, were used.

UIS (2018) argues that the appropriate strategy is to raise awareness of the problems in these harmonised datasets, but to acknowledge that they are valuable in terms of understanding broad global trends. Adjusted country-specific statistics in these datasets ought not be considered in national education debates, however. Put differently, if adjusted statistics display an improvement or deterioration in the quality of schooling for a specific country, this cannot be taken uncritically as a reflection of reality.

Returning to the question of floor effects, how might these effects compromise linking across programmes? The specific question would be the extent to which floor effects might widen the standard errors calculated by Altinok. Altinok's work does not take into account floor effects at all, which is not surprising given how little had been written about these effects. To illustrate, in the case of Argentina, the upper primary proportion proficient in 2013 is 77%, with the 95% confidence interval being 74% to 80%. In theory, it should be possible to estimate how much wider this confidence interval would be if floor effects were taken into account. Crucially, some of the widening of the interval would arise out of a consideration of floor effects in the original programme-specific data, given that even before any linking, there is 'noise' produced by floor effects which is not taken into account. Thereafter, one could consider adapting the portion of the standard error attached to the linking process, in line with Altinok's methodology, to arrive at a new standard error. Obviously, the larger the floor effects, the larger would be the confidence interval. This would be a complex exercise as it would involve examining original item-specific scores and calculating aggregate classical scores, as is done in the current paper. Altinok's work makes use only of the aggregate student-level IRT scores in the assessment microdata, not item-specific scores per student.

To conclude this section, assessing how fit for purpose existing harmonised datasets of proportion proficient are, is still in its infancy. These datasets are useful for illuminating global patterns and trends, though countries are likely to be critical of country-specific patterns considered questionable or clearly incorrect, given the data countries themselves have on proficiency. Interrogation of the methods behind these harmonised datasets should proceed, and in doing so the impact of floor effects is one of several issues that should be considered.

4 Existing work on floor effects

Floor effects have received little explicit attention in the literature. This would in part be because countries most affected by floor effects, namely developing countries, are the least endowed when it comes to assessment data and technical capacity in the area of assessments. Hence the current paper deals to a large extent with establishing the basic features of this phenomenon.

Among the little research that exists is Catts *et al* (2009), who find that floor effects in diagnostic tests seriously limit efforts to target the right assistance to children with learning difficulties in Florida. It would thus not be true to say that floor effects are an exclusively developing country concern. Yet Resch and Isenberg (2014) find that floor (and ceiling) effects do not have serious practical implications in an urban district in the United States.

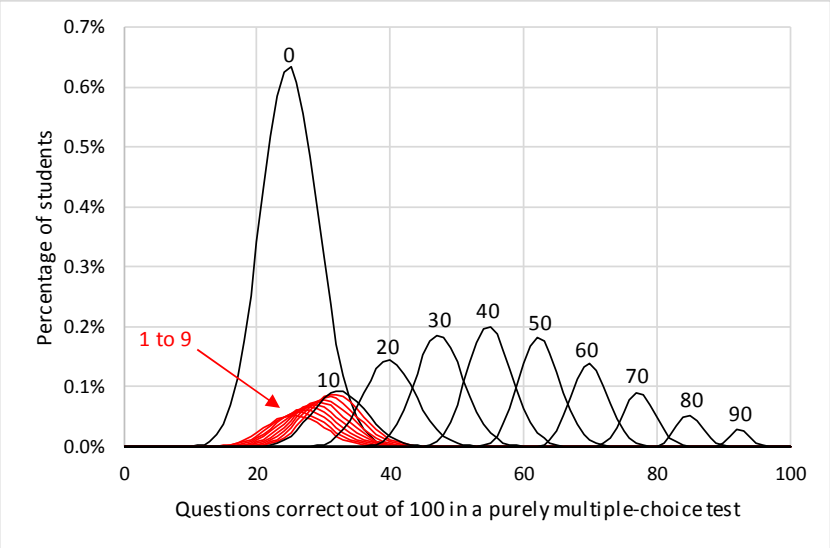
Mullis, Martin and Loveless (2016: 61) discuss floor effects in a twenty-year review of TIMSS. They find that improvements in Grade 4 over the years in TIMSS have tended to be larger

among worse performing students than better performing students, an excellent finding in terms of reducing inequalities. However, the question is asked whether the finding might be a construct brought about by floor effects, as opposed to a real trend. Could the floor have moved upward slightly, thus creating the illusion that the worst performing students perform better in later years? The opinion of the authors is that this is unlikely, given what they know about measurement in TIMSS. The current paper does not attempt to investigate this specific matter empirically. Such an investigation should be possible with the available data, but it would be more ambitious than what is presented below.

Of particular relevance for the current paper is Burton (2001), who takes stock of and augments the theory around the effects of random guessing in multiple choice tests. This, as will be seen below, is an important matter when floor effects in existing testing systems are examined.

Figure 2 explains the essentials of Burton's approach. The graph presents patterns from a theoretically derived set of data, not a dataset of a real students. It is assumed that if the students participated in a test with 100 constructed response items only, so with *no* multiple choice items, on average students would achieve just 35 items correct, and the standard deviation would be 23. These parameters are what one finds in Chile's 2015 Grade 4 TIMSS mathematics test, considering just the *constructed response* part of the test. In that test, there are not exactly 100 items (in fact, as will be explained below, the 'test' is actually different booklets), but on average students obtain 35% of the constructed response items correct. A considerable percentage, 7%, obtain a score of zero. If the assumed students participate in a test consisting of just 100 *multiple choice* questions, what would be the distribution of scores be? This is the question Figure 2 answers. This graph uses as inputs the Chilean mean and standard deviation, and Burton's equation (2) with the assumption of four options per question. If we assume that each multiple choice question has four options, then the tall curve in Figure 2 represents the distribution of scores of those students who would have obtained zero in the constructed response test. The most common score would be 25 correct, and 0.63% of all students would fall into the category of obtaining zero in the constructed response test, but 25 in the multiple choice test. This is due to random guessing – if there are four options per question, on average students who do not understand any questions will obtain 25 out of 100. However, some students who understand no questions will obtain more or less than 25, by chance. Crucially, Burton's approach finds that virtually no students would obtain a score of less than around 11. The red curves in Figure 2 reflect the distributions of scores for students who in a constructed response test would obtain one question right, two questions right, and so on, up to nine questions right. Thereafter, only curves for every multiple of 10 are shown. Clearly, even better performing students benefit from random guessing. For example, the most common multiple choice test score for students obtaining 40 in a constructed response test, is 55.

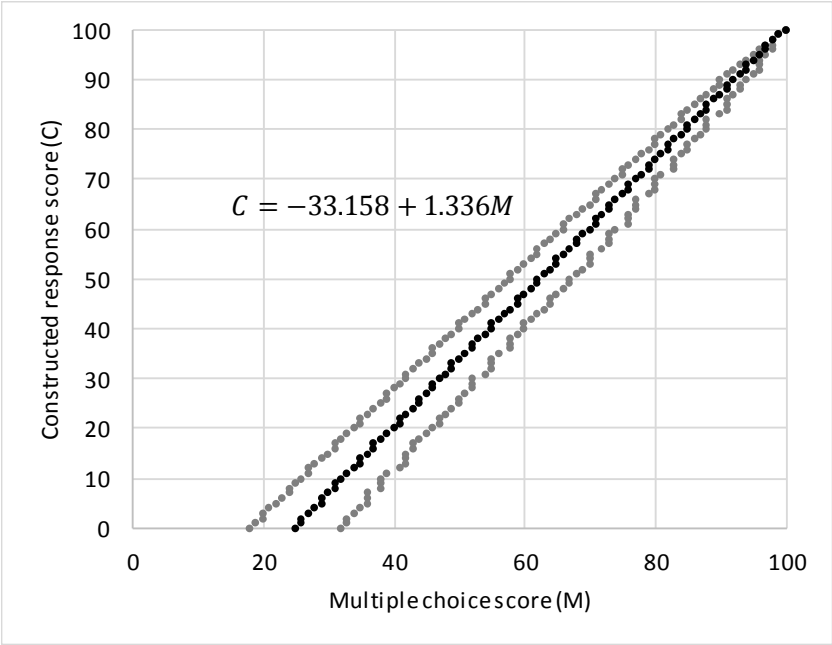
Figure 2: Floor effects in a 35-23 multiple choice test



Note: The value at the peak of each curve indicates the constructed response score of the students represented by the curve.

The values reflected in Figure 2 produce the simple linear relationship seen in Figure 3. To illustrate, the curve for a constructed response (C) score of zero in Figure 2 'peaks' at a multiple choice (M) score of 25. This determines the bottom of the line of black markers in Figure 3. A student who scores 25 in a multiple choice test with four options per question is assumed to score zero in a test consisting entirely of constructed response questions. From the M score we subtract 25. However, if the M score is 60, we subtract 13, giving a C score of 47. A student who achieves an M score of 100 is assumed to score 100 in the C test too. Essentially, the better the student's M score, the less the student is considered to benefit from random guessing, as there is less uncertainty in which to apply random guessing. The grey markers in Figure 3 represent the 95% confidence interval on the horizontal axis – this is discussed below.

Figure 3: Conversion of multiple choice to constructed response score



Note: Each point in the graph represents the peak of a possible Figure 2 curve.

Note that Figure 3 can be used to convert a country's mean score only if data on individual students are available. The conversion must occur initially at the level of the student. Also note that departing from Chile's mean and standard deviation would change Figure 2, but not Figure 3. The equation in Figure 3 is thus indifferent to these two parameters. An additional step not illustrated in the equation is that any value *C* which is less than zero must be raised to zero.

One adaptation of the equation may be necessary. If students provide *no* response to certain multiple choice questions, then obviously they forfeit some of the opportunity of gaining a mark through random guessing. If we apply the equation of Figure 3 without any adaptation, we are likely to *over*-adjust, in other words *under*-estimate what students with null responses would be achieving in the constructed response test. The following equation, devised specifically for the current paper, provides the necessary adjustment.

$$C = \left(-33.158 + 1.336\left(\frac{S}{T-N} \times 100\right)\right) \times \frac{T-N}{T} \quad (1)$$

Here the adjustment illustrated in Figure 3 is applied only to the portion of the test with responses, in other words correct responses (*S*) divided by all items with some response, meaning all items (*T*) minus null responses (*N*). Thereafter, the adjusted percentage correct is adjusted downward, in the case of students with null responses, using items with some response divided by all items. Equation (1) also works where there are no null responses, so where *N* equals zero. To illustrate, a student with *no* null responses who achieves 40 of 60 responses correct, would have a classical score *M* of 66.6 (out of 100), and an adjusted classical score *C* of 55.9. If we assume five omitted or null responses, *C* becomes 58.7. In other words, the downward adjustment of the original 66.6 is smaller as the student took up fewer opportunities to randomly guess answers.

We can use the 95% confidence intervals illustrated by the grey markers in Figure 3 in order to find the overall 95% confidence interval for the mean score referred to earlier, which was said to be 35 in terms of a test consisting entirely of constructed response items. This 35 translates to a mean of 51 in a test consisting entirely of multiple choice items. The 95% confidence interval around this mean would be, using rounded values, 46 to 57. This confidence interval of 11 represents almost 0.5 of a standard deviation – the latter statistic was said to be 23 above. This confidence interval would be due to randomness resulting from random guessing only. It is not due to sample size, which is be a separate matter. In other words, enlarging the sample would not change the random guessing confidence interval. A confidence interval as large as 0.5 standard deviations is enormous. To compare, 95% confidence intervals reported in TIMSS, arising from sampling, seldom exceed 0.15 of a standard deviation, and can be as low as 0.08 (Mullis, Martin, Foy *et al*, 2016: Exhibit 1.1).

To understand how problematic a confidence interval of around 0.5 of a standard deviation is, one can think about its impact on comparisons across countries, and on comparisons of one country's progress over time. A confidence interval of this magnitude would make the rankings of, for instance, SACMEQ countries a bit less clear than they are generally presumed to be. SACMEQ is a relevant example here as all of SACMEQ's tests consist of multiple choice questions only, as will be seen in section 6 below. Moreover, fourteen of fifteen SACMEQ countries displayed classical scores of 51% correct or less in 2007 in mathematics – the exception is Mauritius, which achieved 58%. Yet the biggest problem would not be comparisons across countries, but comparisons for a country over time. Progress over time is almost inevitably slow, according to past trends seen in even those countries with the best education improvement interventions. Roughly, one can think of a 'speed limit' of 0.06 standard deviations a year (Gustafsson, 2014: 134). A testing programme with confidence intervals as large as 0.5 of a standard deviation would not be in a position to monitor progress over time in the way most education authorities would want to do this. To illustrate, a confidence interval

of 0.5 standard deviations would not allow one to reliably gauge a relatively exceptional and impressive improvement of 0.03 standard deviations a year – half of the 'speed limit' – over any period shorter than ten years. Clearly, an education authority would want to know whether improvement interventions were working before ten years had passed.

It is not the intention of the current paper to go further and quantify with greater precision the degree to which floor effects interfere with comparisons, in particular comparisons over time. Any such analysis would have to combine margins of error associated with random guessing with margins of error associated with the sampling approach. The latter would clearly exacerbate the situation. IRT scores would, to a limited extent, improve comparability, but IRT would by no means resolve the problem. Finally, the rough quantification presented here deals with the mean. Of relevance for the SDG monitoring debates is the proportion proficient, not the mean. However, margins of error associated with the former are likely to be as problematic as those associated with the latter.

5 An initial exploration of floor effects using TIMSS Grade 4

5.1 The TIMSS 2015 Grade 4 mathematics data

The 2015 Grade 4 microdata for the subject mathematics, from the TIMSS programme, available online through the TIMSS-PIRLS website[5], offer excellent insights into the matter of floor effects. Since the 2015 wave of TIMSS it has been possible for participating countries to test Grade 4 students using either the regular TIMSS assessments existing for many years, or a new TIMSS Numeracy assessment which is less demanding. The 2015 results from both assessments were expressed in the same metric as the tests could be linked through common items. TIMSS Numeracy was introduced specifically to counteract floor effects that resulted when countries with large proportions of poorly performing students participated in the regular TIMSS testing.

In five countries in 2015, around half of the Grade 4 students took the regular TIMSS test, and the other half TIMSS Numeracy, which allows one to gauge how the two tests function with the same student population. In the five countries, students in the same school were randomly assigned to one of the two tests. It was not a case of better performing students taking the regular TIMSS test. The five countries are: Bahrein, Indonesia, Iran, Kuwait and Morocco. The published national mathematics scores for these five countries are simply the average across two means: the regular TIMSS mean and the TIMSS Numeracy mean. Two additional countries, Jordan and South Africa, participated only in TIMSS Numeracy. One can speculate that these two countries were more certain that in future they would participate in just TIMSS Numeracy, while the other five were less certain about the nature of their future participation. Largely as a result of the introduction of TIMSS Numeracy, the number of countries considered by TIMSS to suffer from reliability problems, through floor effects, in their Grade 4 mathematics results declined from five countries in TIMSS 2011 to two countries in TIMSS 2015. The two were Kuwait, despite its participation in TIMSS Numeracy, and Saudi Arabia, which continued to participate in just the regular TIMSS test in 2015. (Mullis *et* al, 2012: 40; Mullis, Martin, Foy *et al*, 2016: Exhibit 1.1.)

5.2 Comparing classical and IRT scores

Figure 4 below illustrates how the item response theory (IRT) scores used in the official TIMSS reports compare to the underlying classical scores. Classical scores are what teachers would be highly familiar with: percentage correct, or the score obtained by the student divided by the maximum possible score. The classical scores were calculated from the TIMSS data by comparing each student's data with item information tables published by TIMSS. Both graphs

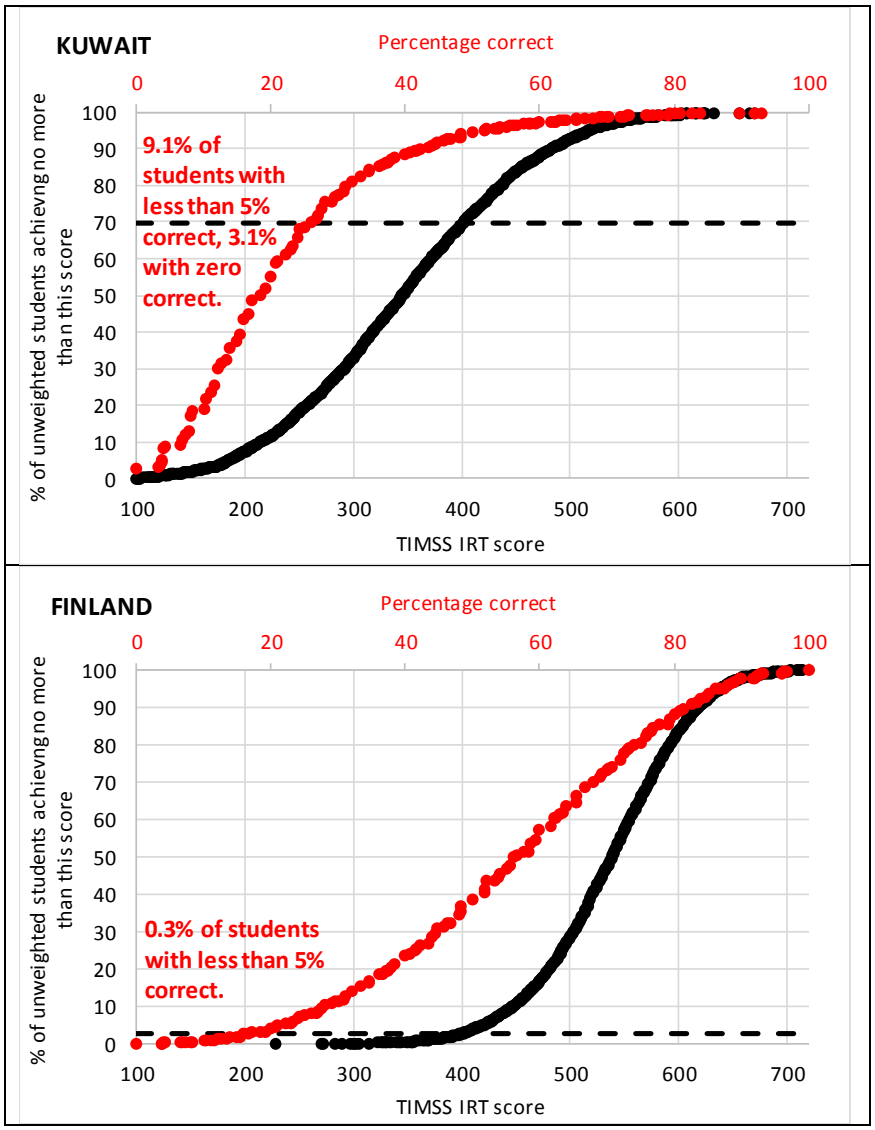
[5] https://timssandpirls.bc.edu.

in Figure 4 reflect regular TIMSS results (not TIMSS Numeracy). IRT assists in reducing floor effects, relative to classical scores, for two reasons. Firstly, IRT scores take into account the difficulty of each item. If two students achieved just two questions correct, but the first student's two questions were more difficult questions than the second student's two questions, then the first student would obtain a higher IRT score, even if they both obtained the same classical score. This largely explains why there is such a large variety of IRT scores for poorly performing students in both Kuwait and Finland – the left-hand ends of the IRT curves are stretched out horizontally. In Kuwait, the worst performing 5% of students, according to the classical scores, had just five different classical scores (0.0, 3.3, 3.4, 3.7 and 3.8 per cent correct). In contrast, the worst performing 5% of students according to the IRT scores display a large range of IRT scores, in fact virtually every student has a unique IRT score in the range of 8 to 183. Clearly, the IRT scores of the worst performing students are spread out rather widely. There is apparently no floor effect.

An important methodological point should be clarified at this stage. Due to the matrix sampling approach used by TIMSS in constructing its tests, meaning essentially the development of different overlapping test booklets of equivalent difficulty, each student is assigned five 'plausible' IRT scores (Martin, Mullis and Hooper, 2017: 4.1). This means special 'plausible value' techniques must be employed when calculating descriptive statistics such as the mean, or the percentage of students passing a performance threshold. To illustrate distributions of the IRT scores, as in Figure 4, one of the five plausible values should be used. Which one is used makes no discernible difference to the graph. However, one should avoid using the average across the five plausible values as this *does* result in a slightly different pattern.

The second reason why IRT scores seem to eliminate floors is that data *from outside the test* are used to refine each student's score. This explains why in Kuwait, the 110 students who obtained a classical score of zero, meaning they got no answers correct, display 110 different IRT scores in the range of 49 to 384 (the student with an IRT score of 8 was a non-zero student obtaining a score of 2 for the multiple choice items – the value 8 was referred to as a minimum previously). The data from outside the test are largely the results from the science test, which is run parallel to the mathematics test. However, even a student's home background data may be used to predict differences across students who obtained a zero classical score (Martin *et al*, 2016: 13.20). If one were to calculate IRT scores *without* using any information from outside the mathematics test, then all students obtaining a classical score of zero would be assigned virtually the same IRT score – there could still be variation across the five plausible values of each student (assuming a matrix sampling approach were followed).

Figure 4: IRT compared to classical scores in regular TIMSS



Note: The red curves refer to classical scores, and should be read against the top horizontal axis. The horizontal dashed lines reflect the percentage of students achieving no more than the TIMSS 'low international benchmark' of 400.

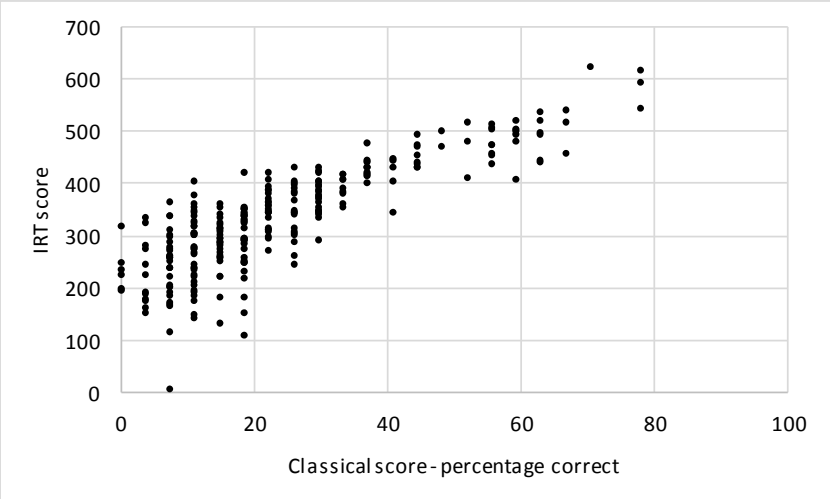
While IRT scores display no serious floor effects in either Kuwait (where one might expect them) or Finland (where one would not), the classical scores in Kuwait do display this. In Kuwait, 9.1% of students obtained less than 5% correct, and 3.1% of students obtained 0% correct.

One practical lesson for policymakers can be noted at this point. One way of ensuring that an assessment differentiates to a greater extent among students at the low end of the achievement continuum, is to use IRT. Capacity to score results using IRT is still weak in many countries, yet investing in this capacity has several benefits, including more informative results as far as worse performing schools and students are concerned. It is of course possible to replicate just some of the TIMSS methods, as opposed to all of them, to reduce complexities. In particular, allowing results from other subjects, or even a student's home background, to influence IRT scores is probably best avoided at an initial stage. At an initial stage, the emphasis should probably fall on using IRT to bring about a better differentiation between students with low

non-zero scores. Even with IRT, students with a zero classical score would not be differentiated at this initial stage.

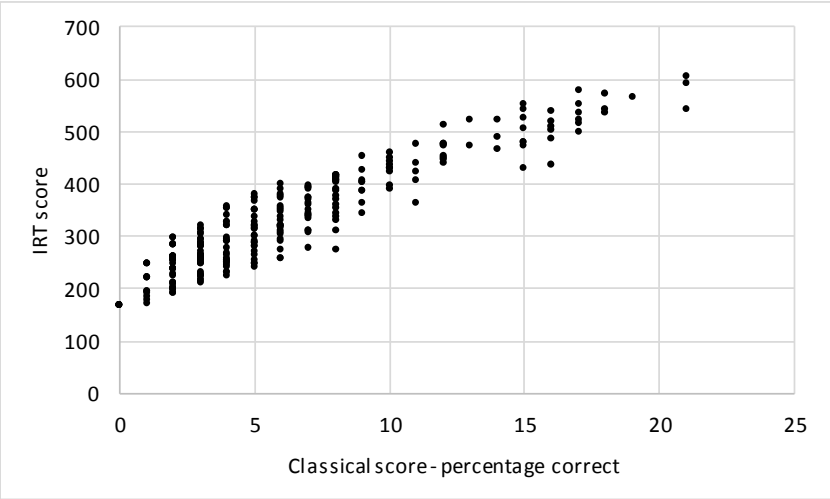
Figure 5 reinforces what has been explained above. Not only do IRT scores assist in differentiating between students at the lowest end of the performance spectrum, they do so across the entire continuum. Figure 5 illustrates the results of only those 264 Kuwait students whose test consisted of Book 1 – this would be one of the fourteen TIMSS Grade 4 test booklets. According to Figure 5, even students achieving around 60% correct would be differentiated from each other according to the level of difficulty of the items they got correct.

Figure 5: IRT compared to classical scores in Kuwait Book 1



But what would Figure 5 look like if a simpler IRT score calculation were employed, which did not make use of data outside the mathematics test? The result would be less differentiation, but still more than enough to justify the use of IRT. Figure 6 illustrates the result of a calculation of IRT scores using only the raw item response data from students taking Book 1. The 264 students would display 245 unique IRT scores. This figure 245, is also the number of combinations of specific items correct in the data.

Figure 6: Recalculated IRT compared to classical scores in Kuwait Book 1

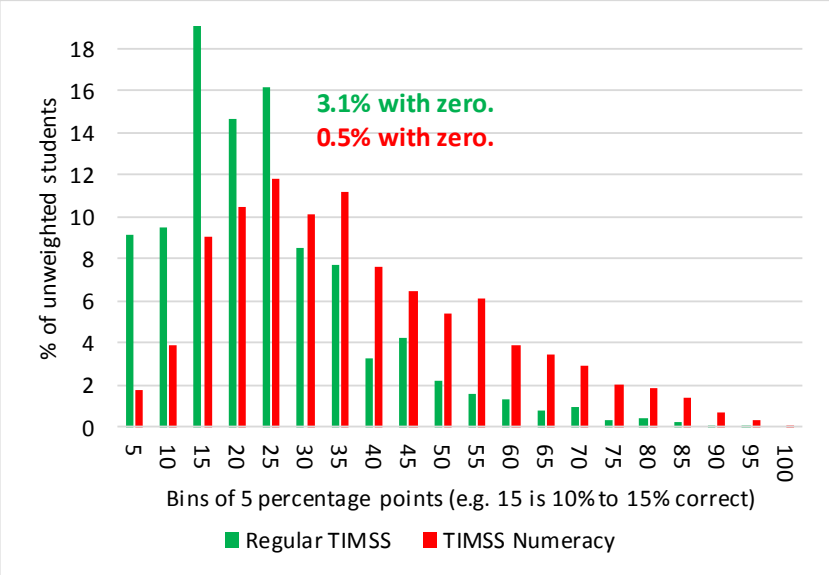


Note: IRT scores were calculated using clogit *in Stata as explained by Jeroen Weesie at https://www.stata.com/support/faqs/statistics/rasch-model. Thereafter, standardisation of values to produce the same mean and standard deviation as for the Figure 6 values was employed.*

5.3 The benefits of the easier TIMSS Numeracy test

Next, the benefits brought about by TIMSS Numeracy are considered. Figure 7 illustrates the distribution of classical scores in regular TIMSS and TIMSS Numeracy, in 2015, for Kuwait. Kuwait was the country with the most serious floor effects insofar as it was the only country participating in TIMSS Numeracy considered to display floor-related reliability problems according to the official TIMSS report. The insertion of easier items into TIMSS Numeracy resulted in a reduction in the percentage of students with a zero classical score from 3.1% (in regular TIMSS) to 0.5%. The floor was thus pushed downward, resulting in fewer students 'under the floor'.

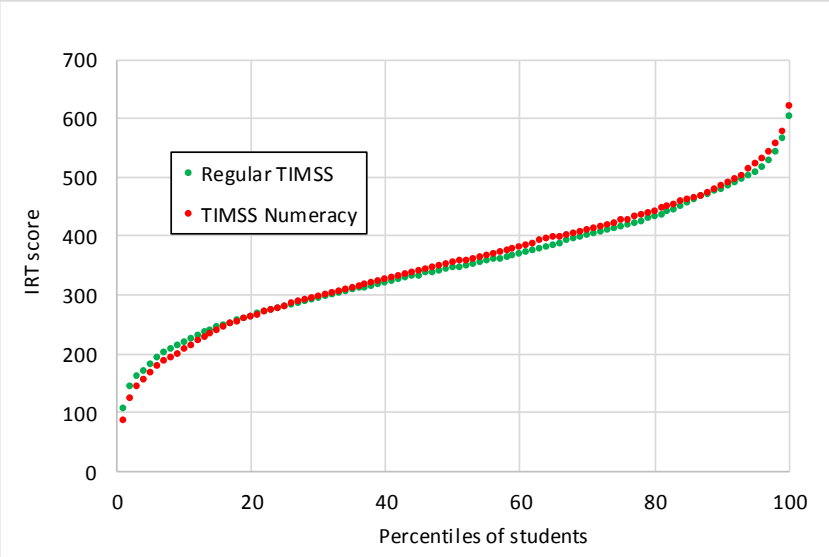
Figure 7: Regular TIMSS compared to TIMSS Numeracy in Kuwait (classical)



A second practical lesson can thus be that an assessment's level of difficulty should be in line with the actual competencies of the student population. Not only does this reduce or eliminate floor effects, it also produces better differentiation between students across the entire continuum. To illustrate, the middle group of students in Kuwait located between the 25th and 75th performance percentiles (meaning Kuwait's students if one excluded the top and bottom quarters), displayed classical scores of between 12% and 28% in regular TIMSS, but between 19% and 47% in TIMSS Numeracy. TIMSS Numeracy was more successful in differentiating between the students, with the ranges being 16 and 28 percentage points in regular TIMSS and TIMSS Numeracy respectively.

What is said in the previous paragraph refers to classical scores. In terms of IRT scores calculated by TIMSS, differences between regular TIMSS and TIMSS Numeracy are barely noticeable. This can be seen in the following graph, where the two curves are barely distinguishable. This is because regular TIMSS, even in the face of problems such as 3.1% of students with a zero classical score, is rather successful at differentiating students even at the low end, in part due to the imputations using external data discussed above. Yet, as will become clearer below, relying on tests set at a level of difficulty which is inappropriately high, is problematic.

Figure 8: Regular TIMSS compared to TIMSS Numeracy in Kuwait (IRT)



5.4 Multiple choice questions and random guessing

The focus now turns to a critical matter: the mix of multiple choice and constructed response items in the assessment. The following table sums up the situation regarding items and books in the two TIMSS assessments. There were fourteen regular TIMSS books, and five for TIMSS Numeracy. Each student responded to questions in just one mathematics book. In regular TIMSS there were slightly more multiple choice questions than constructed response items – 83 against 74. In TIMSS Numeracy, the opposite is found. However, roughly there was in both cases an equal spread across the two question types. Importantly, there were twice as many items per book in TIMSS Numeracy as in regular TIMSS – 51 against 26 on average. This points to an important cost of dealing with floor effects through the insertion of more easy items: the test becomes longer. 157 regular TIMSS items were each repeated across two different regular TIMSS books, and the 107 TIMSS Numeracy items were each repeated across two different TIMSS Numeracy books. The 22 items found in both were repeated across four books, two regular TIMSS books and two TIMSS Numeracy books.

Table 1: Details on TIMSS Grade 4 books and items in 2015

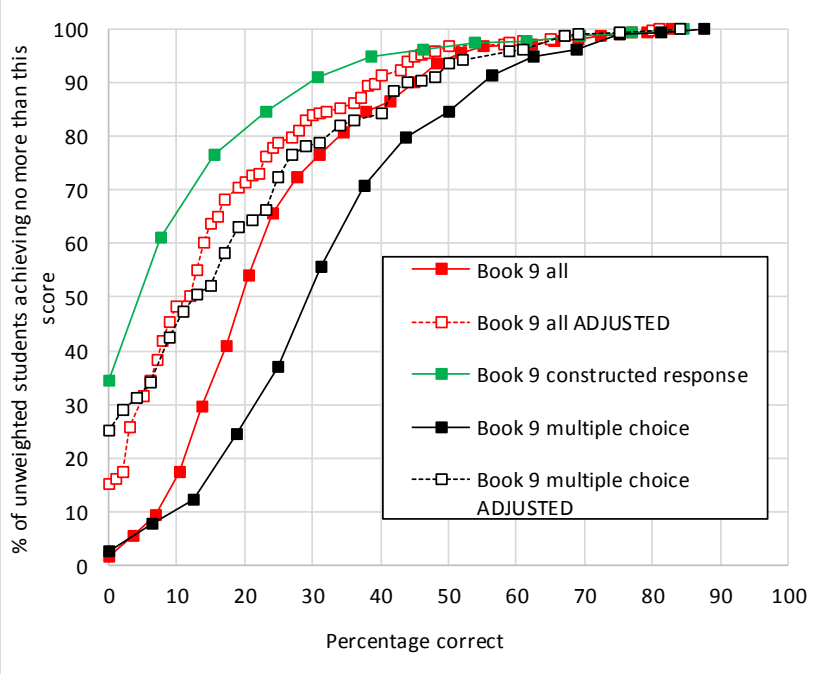
	Regular TIMSS	TIMSS Num-eracy	Both	Total
Books	14	5		19
Items	157	107	22	286
Multiple choice	83	50	11	144
Constructed response	74	57	11	142
Average items per book (including items found in both)	25.6	51.2		

Figure 9 illustrates the distribution of regular TIMSS item-specific responses in Kuwait, for just one of the fourteen books, namely Book 9. Similar patterns would emerge if one of the other thirteen books were analysed. The floor effects are clearly more serious than the preceding analysis would suggest. A whole 34% of students scored zero in the twelve constructed response questions in Book 9. Just 3% of students, however, scored zero in the fifteen multiple response questions. This is clearly due to the random guessing discussed in section 4. Overall, just 2% of students obtain a score of zero overall, meaning zero for *both* the constructed response and multiple choice parts of the test.

But what if the multiple choice score for each student is adjusted, using equation (1)? And if one adjusts the score, how important is it to differentiate between an incorrect response and no response at all? This differentiation, it turns out, is important as null responses in the TIMSS data are not negligible. In Kuwait's regular TIMSS, null responses account for 13% of all multiple choice item responses, the figure for TIMSS Numeracy also being high, at 10%. Kuwait's figures are particularly high. The average across the countries covered in Table 2 below is 6% in regular TIMSS, and 5% in TIMSS Numeracy[6]. An important point for understanding null responses is the following. Null responses as a percentage of all incorrect responses is *lower* for better performing students, across all countries. Thus, the situation is not one where better performing students are *more* inclined to return a null response where they are uncertain about the correct response. This is also means that worse performing students are more likely to 'benefit' from the element of equation (1) which takes into account a student's propensity *not* to make use of random guessing.

If one converts multiple choice scores to scores one assumes would have appeared without random guessing in an equally difficult set of constructed response questions, one obtains the 'Book 9 multiple choice ADJUSTED' curve in Figure 9. This curve points to a whole 25% of students achieving zero. This curve moreover produces an average multiple choice score across all Book 9 students of 19%. The figure would have been 16% if null responses had not been treated differently. The constructed response curve in Figure 9 produces a not very different average score of 13%. This closeness is by design. In the TIMSS documentation, item parameters, which are measures of item difficulty, are on average similar if one compares parameters for constructed response items and parameters, with a guessing adjustment, for multiple choice items (Appendix 13A of Martin *et al*, 2016).

Figure 9: Floor effects before and after random guessing adjustments



Source: Data for Kuwait corresponding to Book 9 of regular TIMSS.
Note: Adjusted scores were rounded to the nearest whole percentage.

The 'Book 9 all ADJUSTED' curve indicates that 15% of students achieved zero for the test as a whole after the adjustment of the multiple choice part of the test. This would be line with the TIMSS report's explanation of why reservations were expressed for Kuwait and Saudi Arabia:

[6] TIMSS distinguishes between 'omitted' and 'not reached'. Both these categories are included in these statistics.

'the percentage of students with achievement too low for estimation exceeds 15% but does not exceed 25%' (Mullis *et al*, 2016: Exhibit 1.1). Of course, Kuwait also participated in TIMSS Numeracy and, as will be seen below, here floor effects were negligible.

Table 2 provides 'under the floor' percentages for eleven countries, including four participating only in regular TIMSS – one relatively weak performer (Saudi Arabia), two strong performers (Finland and Singapore) and one middling performer (Chile). In regular TIMSS, only Finland and Singapore display what can be considered completely non-worrying levels of students 'under the floor', or students with a zero classical score, after the adjustments for guessing discussed above are applied. One can expect that even in the best schooling systems, a small number of students would obtain zero simply because they were not feeling well, or because they willingly 'sabotaged' the test. Importantly, TIMSS typically excludes intellectually disabled students from the testing, as well as students whose school language is not catered for in the tests (Martin *et al*, 2016: p. 3.6). It would be difficult to set a hard threshold for an acceptable extent to which students score zero. Yet situations such as that of Saudi Arabia, where 12% of students achieve zero, after guessing adjustments, are obviously problematic. Saudi Arabia should clearly participate in the easier TIMSS Numeracy, which the country did not do in 2015.

Table 2: Percentage of students 'under the floor' across 11 countries

	No guessing adjustment			With adjustment	
	Constr. response	Multiple choice	Overall	Multiple choice	Overall
Regular TIMSS					
Bahrain	9.9	1.2	0.5	18.0	4.3
Chile*	6.9	1.1	0.4	13.1	2.4
Finland*	1.4	0.3	0.1	3.9	0.4
Indonesia	21.2	1.3	0.7	25.7	9.3
Iran	13.3	1.6	1.0	17.9	5.4
Kuwait	33.6	4.7	3.1	36.5	17.5
Morocco	27.5	2.4	1.1	32.3	12.9
Saudi Arabia*	25.0	2.5	1.4	31.3	11.8
Singapore*	0.7	0.1	0.0	2.2	0.3
TIMSS Numeracy					
Bahrain	0.7	0.1	0.0	3.7	0.3
Indonesia	1.1	0.0	0.0	6.6	0.6
Iran	1.0	0.1	0.1	2.9	0.3
Jordan‡	4.5	0.6	0.4	10.0	1.9
Kuwait	4.8	0.9	0.5	13.2	2.4
Morocco	2.4	0.3	0.1	13.7	1.1
South Africa‡	1.1	0.2	0.0	13.4	0.5

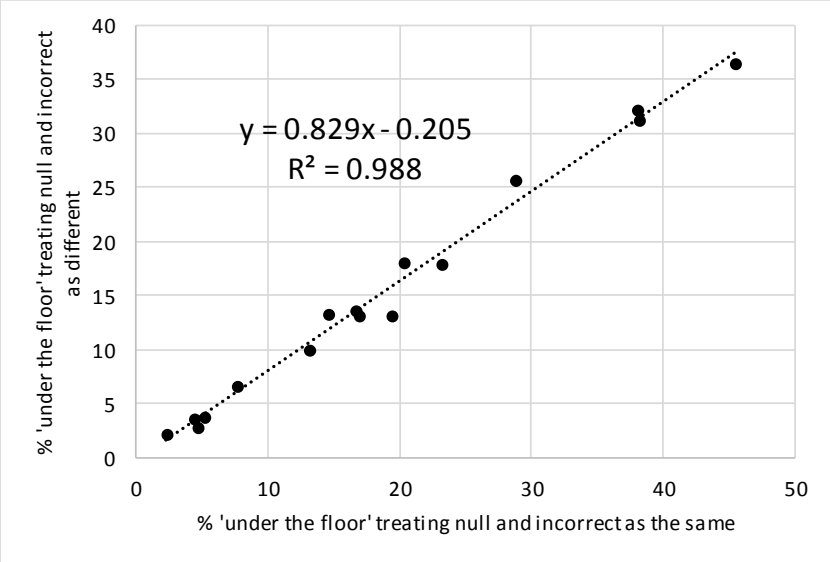
*Note: ‡ means participated only in TIMSS Numeracy, * only in regular TIMSS.*

In TIMSS Numeracy, the percentages of students with a classical score of zero, after adjustments, are all quite low, though arguably still a bit worrying for Kuwait, at 2.4%. It is moreover reassuring that students with zero in the constructed response items is so much lower in TIMSS Numeracy than regular TIMSS – for instance Kuwait's 4.8% against 33.6%.

A critical question for the current paper, which includes analysis of datasets where it is not possible to differentiate null from incorrect, is whether the TIMSS data reveal patterns which can inform, in a rough sense, multiple choice score adjustments when using these other datasets. Figure 10 points to a remarkably predictable relationship in TIMSS between, firstly, 'under the floor' percentages which do *not* differentiate between null and incorrect, and ones which *do* make this differentiation. For no differentiation, one simply applies the equation seen in Figure 3 above. Each point in Figure 10 represents a row from Table 2, and the vertical axis refers to the last column of that table. The horizontal axis refers to what the figure in the last column

would have been if null had not been differentiated from incorrect. Clearly, not making this differentiation results in a higher 'under the floor' statistic. The relationship shown in Figure 10 will inform the discussion below where data which only allow one to calculate the statistic represented by the horizontal axis are used.

Figure 10: Effects of differentiating null from incorrect



As one might expect, if one breaks the figures in the last column of Table 2 down by level of socio-economic disadvantage, one finds larger proportions of students 'under the floor' among the more disadvantaged. This can be seen in Table 3, which uses the availability of books in the home, a typical indicator of home background disadvantage, to disaggregate. In every country, there is a large difference.

Table 3: Percentage of students 'under the floor' by socio-economic status

		% of students 'under the floor' with adjustment		
	% of students from extremely 'book poor' households	% from 'book poor' households (A)	% from households with 11 or more books (B)	A / B
Regular TIMSS				
Bahrain	21	6.1	4.1	1.46
Chile*	32	4.3	2.8	1.53
Finland*	5	1.2	0.5	2.54
Indonesia	43	13.2	7.8	1.70
Iran	41	10.3	4.8	2.15
Kuwait	32	23.3	18.7	1.24
Morocco	56	16.1	12.9	1.25
Saudi Arabia*	36	13.9	12.9	1.07
Singapore*	10	1.2	0.1	9.86
TIMSS Numeracy				
Bahrain	21	0.6	0.2	2.86
Indonesia	44	0.8	0.3	2.90
Iran	40	1.0	0.3	3.65
Jordan‡	41	3.4	1.8	1.87
Kuwait	31	3.5	2.6	1.32
Morocco	57	1.5	1.2	1.26
South Africa‡	47	0.8	0.3	2.44

Note: The first column of statistics uses the TIMSS student weights.

A breakdown like the one in Table 3 by gender would reflect better performance among girls than boys, which would be in line with the fact that average scores among girls are higher, as shown in the official TIMSS reports, not just for Grade 4, but even Grade 8.

If one applies the kind of analysis seen in Table 2 to explore *ceiling* effects, or the prevalence of obtaining a classical score of 100%, it emerges that such effects do in fact exist. Singapore stands out in this regard, with 5.1% of students scoring 100%. But this is rare, even in highly performing countries. In Finland, just 0.3% of students obtained 100% in Grade 4 TIMSS 2015 mathematics.

5.5 Floor effects and proportion proficient statistics

A key question is whether the percentage of students 'under the floor', in the sense of the classical scores, approximates, or even exceeds, the percentage of students said *not* to have reached a minimum acceptable level of proficiency. This is a crucial matter in terms of reporting against SDG 4.1.1. To take an extreme illustration, if IRT scores lead to the conclusion that 30% of students of a country perform *below* the minimum benchmark, but 40% of students obtain a classical score of zero, then clearly the first statistic would be questionable. One would then be classifying certain students as having reached a benchmark, and others as not having reached a benchmark, within a group of students who all obtained zero in the test. Fortunately, at least in TIMSS, this situation does not arise. The 'low international benchmark' of TIMSS, an IRT score 400, has been proposed as one possible threshold for reporting against SDG 4.1.1 and this is used in Table 4 (UNESCO, 2018: 62). Clearly, no country gets close to the undesirable extreme. For instance, in the case of Kuwait in regular TIMSS, 67% of students perform *below* the 400 benchmark (see 33% in Table 4), but from the previous table we see that 21% obtained a score of zero. Those below the benchmark are a mix of students with zero, and with non-zero low scores. If one were to instead use the 'intermediate' benchmark of 475, floor effects pose an even smaller measurement risk, as even more non-zero students would fall below the benchmark.

Table 4: Percentage of students reaching TIMSS benchmarks

	Reaching low TIMSS benchmark of 400		Reaching intermediate TIMSS benchmark of 475	
	%	Conf. Int.	%	Conf. Int.
Kuwait	33	30-36	12	10-14
South Africa	39	36-42	17	15-19
Morocco	41	38-44	17	15-19
Saudi Arabia	43	40-46	16	14-18
Indonesia	50	46-54	20	18-22
Jordan	50	48-52	21	19-23
Iran	65	62-68	36	34-38
Bahrain	72	70-74	41	39-43
Chile	78	75-81	42	39-45
Finland	97	96-98	82	80-84
Singapore	99	98-100	93	91-95

Source: Means and standard errors used to calculate 95% confidence intervals from Mullis et al *(2016: Exhibit 2.2).*

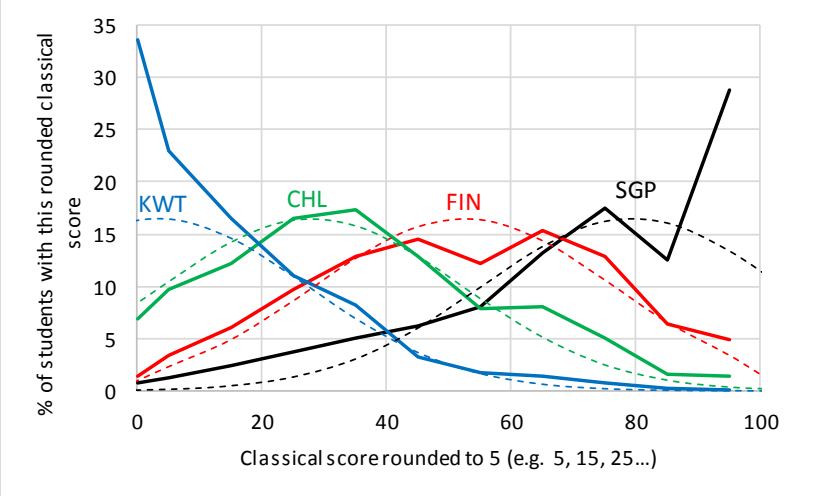
Kuwait's graph in Figure 4 indicates that in regular TIMSS, the classical score threshold for the low international benchmark of 400 TIMSS points, is around 26% correct (reading on the top horizontal axis at the point where the red and black dashed curves intersect). If one performs a similar analysis using Kuwait's TIMSS Numeracy, one finds that the 400 benchmark corresponds to around 43% correct. Neither of these two percentage scores are unreasonably low for a low benchmark, further underscoring the point that floor effects are not a serious

impediment in TIMSS, even in regular TIMSS, to measuring the percentage of students attaining a basic level of proficiency.

5.6 Floor effects and means

While floor effects do not complicate in any direct way reporting against SDG 4.1.1, at least when TIMSS data are used, floor effects are likely to be problematic when *mean* scores are compared. Figure 11 below illustrates the problem. Solid curves represent the distributions of regular TIMSS results, using just the constructed response items (the aim here was to avoid bringing in the complexity of guessing adjustments for multiple choice responses). The floor effects in Kuwait, but even Chile, are visible, as is the ceiling effect in Singapore. Of the countries represented, Finland displays the least floor or ceiling effects, and is therefore used as an anchor for an adjustment process where these effects are accounted for in the other three countries. Finland's dashed line is a normal curve reflecting the actual mean and standard deviation for Finland. This curve is replicated, and placed to the left and right of Finland, as close as possible to the actual curves of the other countries. It is assumed that the curve will always have Finland's original standard deviation, of 24. Clearly, large parts of the normal curve for Kuwait fall to left of classical score zero. Similarly, the Singapore's normal curve extends beyond the right of the graph. We can think of some students in Kuwait (or Chile) scoring below zero, in an imaginary zone where much easier questions are asked. Where the actual curve for Kuwait produces a mean of 14% correct, the normal curve produces a mean of just 3% as imagined scores below zero are taken into account. The generation of the normal curves is somewhat crude, but the approach is good enough to explore what the general impact is of floor (and ceiling) effects on mean values in a constructed response test.

Figure 11: Floor and ceiling adjustments for constructed response items

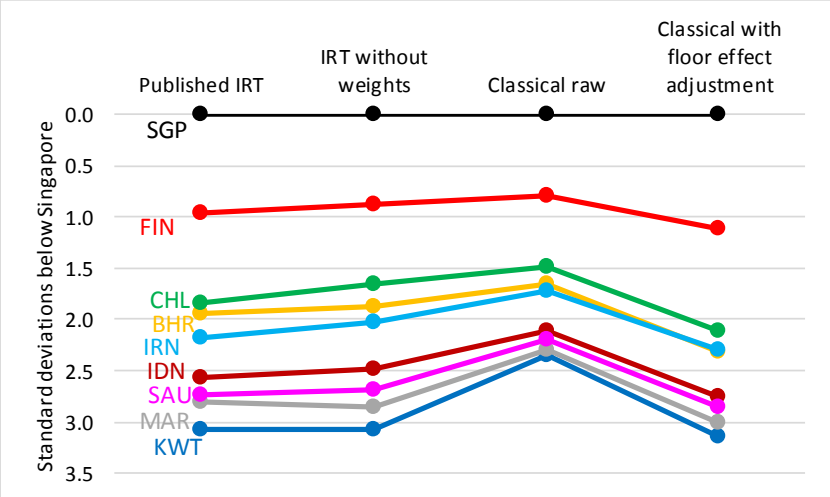


Note: Values 1% to 9% correct were rounded to 5%, 10% to 19% was rounded to 15%, and so on. Exactly 0% correct was maintained as a separate category for actual distributions. Normal curves were moved horizontally to the point where they were closest to the actual curve, but considering only the categories 25% to 75%, or categories which are most outside any floor or ceiling. Closeness was determined by looking at the sum of the vertical gaps at the six categories.

Figure 12 below illustrates the rankings and distributions of country means for nine regular TIMSS countries examined earlier, using four different methods. In the first method, means and standard deviations as published in the official TIMSS reports were used. The only thing that is different about the second method, is that student weights were not used, as the analysis of classical scores presented here does not use weights. Evidently, not using the weights hardly changes the picture. In the third method, the means of the actual classical scores, using only

constructed response items, are reflected. In the fourth, adjusted versions of the third column appear, using normal curves as in Figure 11. The vertical placement of countries is in terms of Singapore standard deviation distances from the Singapore mean (the Singapore standard deviations are 86 for the IRT scores, and 24.2 for the raw classical scores).

Figure 12: Floor and ceiling adjustments for constructed response items



The difference between the third and fourth method results is noteworthy. Not taking into account floor and ceiling effects when considering classical scores results in a compression of differences – Kuwait performs at 2.3 standard deviations below Singapore (third column) against a difference of 3.1 when the adjustment is applied (fourth column). However, and importantly, the fourth column does not look very different from the first (or second) column. What this means is that the IRT scoring in TIMSS not only produces more differentiation among weaker students, it also helps to produce apparently accurate distances between the performances of different countries. This is necessary if countries are to set targets. If, say, Indonesia is to set as a target the mathematics achievement at Chile in 2015, then an accurate idea of the gap between the point of departure and the target in terms of standard deviations will help determine whether the target is feasible (UIS, 2018: 30).

6 Floor effects in SACMEQ and LLECE

6.1 The structure of the tests and the data

SACMEQ and LLECE, focussing on Southern and East Africa and Latin America respectively, are two of three major region-specific international assessment programmes with a history of at least 20 years. The third is PASEC[7], which focusses on Francophone Africa.

International SACMEQ microdata, from the 2000 and 2007 runs of the programme, have been made available widely to researchers through the SACMEQ office. An international version of the 2013 data is not available, though each Ministry of Education in each country has been conducting research using its national data. LLECE microdata are readily available through the LLECE website[8]. For the analysis presented below, 2007 SACMEQ and 2013 LLECE data were used, the name of the latter, and most recent, run of LLECE being TERCE[9]. All grades

[7] Le Programme d'analyse des systèmes éducatifs de la CONFEMEN (Programme for the Analysis of Educational Systems).
[8] http://www.unesco.org/new/en/santiago/education/education-assessment-llece.
[9] Tercer Estudio Regional Comparativo y Explicativo (Third Regional Comparative and Explanatory Study).

and two subjects covered by the two programmes were analysed: reading and mathematics, and Grade 6 and (in the case of LLECE) Grade 3. LLECE data generated from Grade 6 science tests were not analysed. A key step that preceded the analysis presented below was the successful replication of published means per country, for both SACMEQ and LLECE. This provided assurance that weights were being properly employed, and that there were no students missing from the data.

LLECE 2013 produced substantial technical documentation, in Spanish. In contrast, there is virtually no technical documentation for SACMEQ 2007. To some extent, the user of the 2007 data can turn to the earlier technical documentation from SACMEQ 2000, which was extensive, though with noteworthy gaps, in particular in relation to the IRT processes followed. In many respects, LLECE is technically more advanced than SACMEQ. What makes LLECE particularly interesting from a SACMEQ perspective is that, like SACMEQ, it covers countries with limited technical capacity at the national level, and many schools with poor quality teaching and learning. Yet, as will be seen below, LLECE turns out to suffer from particularly serious floor effects. Both programmes have played an enormous role in elevating the understanding of and emphasis on learning outcomes in the policy debates in their respective regions. They have also provided local researchers with vital data, and have facilitated a large number of research papers.

Table 5 presents the basic features of the testing in the two programmes, using for illustration purposes the SACMEQ reading test and LLECE Grade 3 reading test. SACMEQ has not followed a matrix sampling approach. Each student wrote the same reading test in 2007, when most items from the 2000 test were repeated. LLECE did employ matrix sampling, in Grade 3 reading through different combinations of six test books. Each student was given two books, containing around 44 items in total. Crucially, the two reading tests described in Table 5 were based entirely on four-option multiple choice questions, which would increase the opportunities for random guessing, relative to TIMSS, where multiple choice items account for only around half of items. As is explained below, in LLECE there was a limited number of constructed response items in mathematics, as well as a small separate constructed response writing test for both grades.

Table 5: Details on books and items in SACMEQ and LLECE

	SACMEQ 2007 reading	LLECE 2013 Grade 3 reading
Books	1	6
Items	55	65
Multiple choice	55	65
Constructed response	0	0
Average items per book	n/a	22

Source: Analysis of the two sets of microdata, also UNESCO (2016: 64)[10].

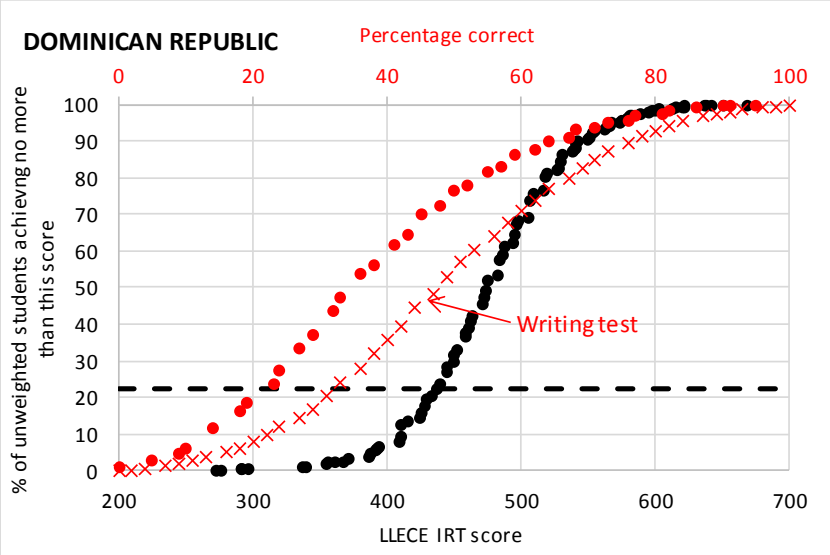
6.2 Comparing classical and IRT scores

Before adjustments for random guessing, there are virtually no students with a classical score of zero in either SACMEQ or LLECE. However, classical scores tend to be lower in LLECE than in SACMEQ and this brings about serious floor effects, though SACMEQ is not immune

[10] There were 55 items in SACMEQ reading for countries writing the test in English. Though items were similar in the three samples where non-English tests were used, some of the 55 were clearly discarded in calculating the official overall scores: 3 in Tanzania and Zanzibar, 7 in Mozambique. These exclusions were also applied in the calculations performed for this paper. Similar issues apply in SACMEQ mathematics.

in this regard. Figure 13 represents the situation in Grade 3 reading in Dominican Republic, easily the worst performer across all tests in LLECE. Only 1% of students achieved zero in the test. The distribution of classical scores in a separate writing test, discussed below, is also illustrated.

Figure 13: IRT compared to classical scores in Dominican Republic (LLECE)

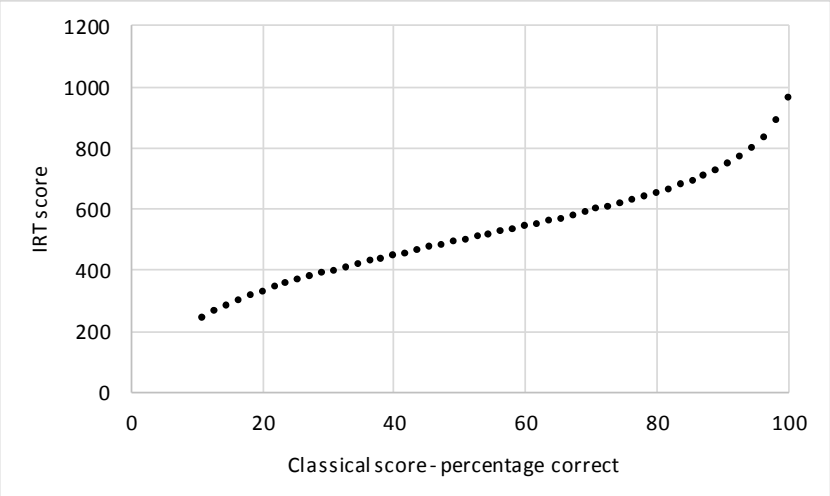


Note: Curves represent scores in Grade 3 reading in TERCE. Red curves should be read against the top axis, the black curve against the bottom axis. The horizontal dashed line reflects the percentage of students achieving no more than an IRT score of 439, the threshold used by LLECE to determine whether a student performed below a very basic level of proficiency. (This 439 was established by examining the microdata and figures in UNESCO [2014: 24]).

Patterns in the LLECE data suggest there is nothing fundamentally wrong with the system of scoring in this programme. SACMEQ, on the other hand, displays odd patterns. As shown in Figure 14 below, each classical score corresponds to just one IRT score. One would have expected a pattern like the one in Figure 5 above. For instance, in Namibia every student obtaining a classical score of 35% also obtained an IRT score of 419. Thus a key benefit of IRT scoring, namely more differentiation across students depending on how well students fared in difficult and less difficult items, was forfeited. There is no obvious motivation for having this pattern in the SACMEQ data, a pattern also found in the earlier 2000 data, and in both subjects[11]. In fact, there are many instances where national assessment programmes *appear* to be making use of IRT scoring, but on closer inspection this technique is being applied incompletely, or incorrectly. All this underlines the importance of developing capacity, including at the national level, in the proper use of IRT.

[11] Further discussion of this problem in Crouch and Gustafsson (2018).

Figure 14: Classical and IRT in Namibia's 2007 SACMEQ reading test



While the pattern seen in Figure 14 is curious, it does not substantially affect the analysis of floor effects which follows.

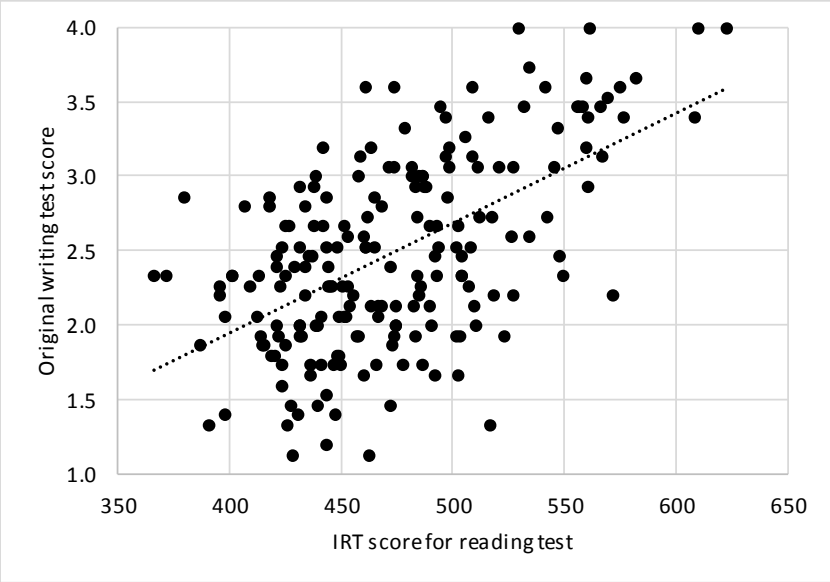
6.3 Multiple choice questions and random guessing

The separate LLECE writing test involved no multiple choice questions and was marked using a marking rubric. Students taking part in the writing test were the students who participated in the other tests. To illustrate, for Grade 3 students were given 45 minutes to write a letter to a friend (UNESCO, 2016: 44). This test appears to present an opportunity to gauge even very basic skills in the absence of the 'noise' produced by random guessing – according to Figure 13, even in Dominican Republic virtually no Grade 3 students scored zero in the writing test[12].

LLECE has apparently not published any results for the writing tests. This would in part be due to comparability problems between the 2013 tests and earlier 2006 writing tests (UNESCO, 2014: 54). However, a further reason is likely to be the very poor correlation between the reading and writing scores. This can be seen in Figure 15. Across all LLECE countries, the Grade 3 correlation was low, around 0.50, between the writing classical scores and the reading scores (whether one used classical or IRT scores for the latter made virtually no difference). The correlation in Grade 6 was only slightly better, at around 0.60 across all countries.

[12] The score range in the test was from 1 to 4. This was changed to a range of 0 to 3 in producing Figure 13, as 1 in the original scoring clearly represents zero.

Figure 15: Reading against writing test score correlations in Dominican Republic

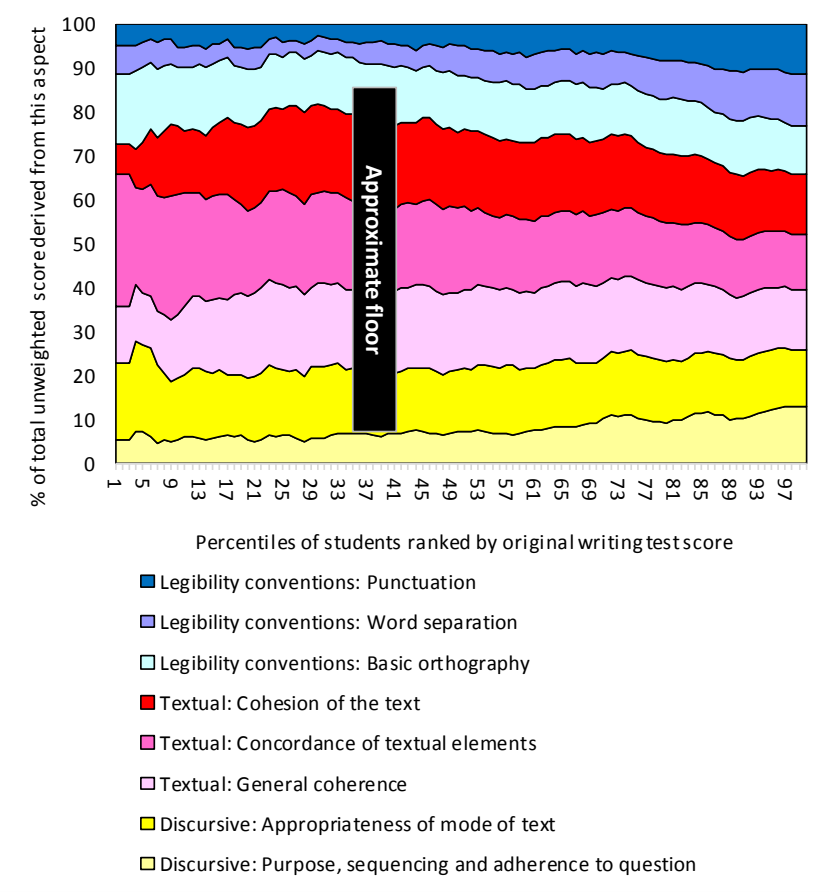


Note: The graph represents 200 randomly selected Grade 3 students.

The LLECE writing test thus offers important lessons around the difficulty of comparable scoring in this type of constructed response test. The reading test scores can be considered relatively reliable – their correlation with the mathematics scores was around 0.70 within all countries (in Grade 3, but also Grade 6). (In SACMEQ, the correlation between mathematics and reading IRT scores is as low as 0.55 in Zambia, though a high 0.83 in Mauritius.)

Even with its problems, there are patterns emerging from the Grade 3 writing test which are interesting in terms of understanding how to assess young children with a weak educational base. Figure 16 illustrates, for Dominican Republic, in which aspects of the writing test students were weaker and stronger. Students with low scores performed particularly poorly in punctuation and word separation. In fact, it is the thickness of these two segments which changes most across the graph. This suggests that one relatively simple remedy for more disadvantaged students is a stronger focus on these basic elements of writing. Weaknesses in 'Purpose, sequencing and adherence to question' in the 'discursive' dimension – this segment of the graph also narrows noticeably in the left-hand side of the graph – would be more difficult to remedy as these are likely to rest on poor reading skills, specifically weaknesses in terms of understanding the test question.

Figure 16: Grade 3 writing competencies in Dominican Republic



Note: Own translation of terms from the original Spanish. In the original marking rules, the overall 'Legibility' score counted for 20% of the total score, and the overall 'Discursive' and 'Textual' scores each counted for 40%. Here these weights are not used as they are not necessary for the points being made. The 'approximate floor' is set at 37%, in line with Table 7 below.

There is of course an extensive literature on what early writing and reading disadvantages consist of. What makes the LLECE reading test exceptional and particularly interesting is that it was implemented on a large scale, and in part with the intention of producing reliable comparisons across countries.

Table 6 produces 'under the floor' statistics using SACMEQ, along the lines of the TIMSS-based statistics appearing in Table 2 (and Table 4) above. The weights in the data are not used here as the focus is just on tested students. In SACMEQ, but also in LLECE Grade 6, as will be seen below, floor effects are far greater in mathematics than in reading. For instance, 37% of students in Zambia obtain zero, after controlling for random guessing, against a figure of 10% for reading. This suggests mathematics tests in particular are more difficult than they should be. Put differently, the mathematics tests do not include enough easy items to allow for sufficient differentiation between the most disadvantaged. That performance is weak in absolute terms in many SACMEQ countries is seen from the estimates of students *not* reaching a minimum threshold: as high as 67% for Zambia in mathematics, and a still high 44% in reading. These figures use as a performance threshold the minimum for the 'basic reading' and 'basic numeracy' category – below this one finds two levels, for instance 'pre-reading' and 'emergent reading'. This use would be in line with what national and international reports have put forward (Moloi and Chetty, 2011; UIS, 2017b).

Key is the fact that in no country is the percentage of students below the threshold – for instance 67% in Zambia's mathematics – *lower* than the percentage of students 'under the floor' – 37% in Zambia. Floor effects do therefore not interfere with proportion proficient statistics. There is no country where the *proficient* include students with a classical score of zero (after adjustments). Had this occurred, and we shall see this *does* occur in LLECE, one would essentially be separating students with the same zero score into non-proficient and proficient groups.

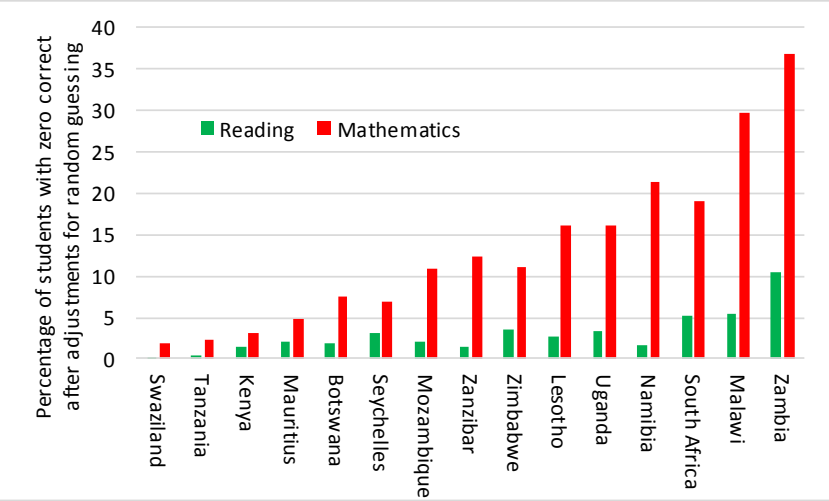
Table 6: SACMEQ 2007 floors and proportions below proficient

	Reading			Mathematics		
	% zero no guessing adjustment	% zero with adjustment	% *not* reaching basic SACMEQ benchmark	% zero no guessing adjustment	% zero with adjustment	% *not* reaching basic SACMEQ benchmark
Botswana	0	2	11	0	7	22
Kenya	0	2	8	0	3	11
Lesotho	0	3	21	0	17	42
Malawi	0	6	37	0	30	60
Mauritius	0	2	11	0	5	11
Mozambique	0	2	22	1	10	33
Namibia	0	2	14	0	20	48
Seychelles	0	3	12	0	7	18
South Africa	0	5	27	0	18	40
Swaziland	0	0	1	0	2	9
Tanzania	0	0	3	0	2	13
Uganda	0	3	20	0	17	39
Zambia	0	10	44	0	37	67
Zanzibar	0	2	9	0	12	33
Zimbabwe	0	4	19	0	12	27

The much larger floor effects in mathematics, compared to reading, are related to much lower classical scores in mathematics. Though IRT scores are set to provide similar magnitudes of performance, for instance 434 in reading and 435 in mathematics in Zambia (Makuwa, 2010), it is in the classical scores that the greater difficulty of the mathematics test becomes clear. In Zambia, for instance, the average classical scores were 38% in reading and 29% in mathematics.

The '% zero with adjustment' figures for SACMEQ are illustrated in Figure 17 below. Here weights in the data *have* been used as the intention is to represent floor effects in the population. The weights do not change the figures seen in Table 6 substantially.

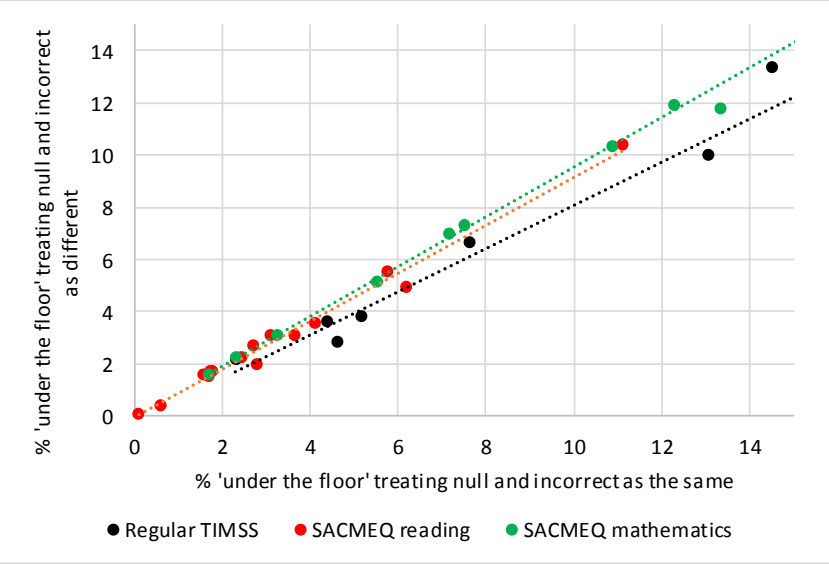
Figure 17: Floor effects in 2007 SACMEQ Grade 6



Note: Countries are sorted by the average across the two subjects. Pupil weights in the data have been used.

SACMEQ data, unlike the LLECE data, do allow one to distinguish between an incorrect response and a null response. This means that for SACMEQ equation (1) could be and was applied in producing the figures in Table 6. Figure 18 illustrates the difference between using the partial adjustment (equation in Figure 3) and the full adjustment (equation (1)). The patterns using the SACMEQ data are in fact similar to those using the regular TIMSS data. This would be because the patterns of missing responses would be similar across the two programmes.

Figure 18: Effects of differentiating null from incorrect (II)



Note: The TIMSS trendline is the same as the one in Figure 10.

As discussed above, floor effects do not seem to interfere with proportion proficient statistics in the case of SACMEQ. But this is when one uses SACMEQ's own performance thresholds. But what if another threshold were used? Specifically, a *lower* threshold would result in fewer non-proficient students, which would increase the chances of the percentage of students 'under the floor' exceeding the percentage who are non-proficient. In fact, no-one has proposed a lower threshold for SACMEQ countries than the one applied by SACMEQ. Of Altinok's (2017) two

thresholds, the lower one representing a 'basic proficiency level', is the same as the one applied by SACMEQ[13].

The attention now turns to LLECE. We see from Table 7 below for LLECE Grade 3 mathematics that the calculation of students 'under the floor' involves separating multiple choice questions from constructed response questions, as in TIMSS. This is because a few constructed response questions were included, two per student, or one in each of six books[14]. The same applies to Grade 6 mathematics, though the LLECE reading tests consisted entirely of multiple choice questions. It is clear from Table 7 that there is a serious floor effects problem in LLECE. In every country and both subjects, the percentage with zero after the adjustment exceeds the percentage of students considered non-proficient. Again, the proficiency threshold used is that of the programme itself. LLECE's reports refer to four levels of performance, from a basic level I to an advanced level IV, and then a fifth level, below level I (UNESCO, 2014: 18). Table 7 reflects students at this very lowest level.

Table 7: LLECE 2013 floors and proportions below proficient in Grade 3

	Reading			Mathematics			
	% zero no guessing adjustment	% zero with adjustment	% *not* reaching basic LLECE benchmark	% zero no guessing adjustment	% zero with adjustment: just multiple choice	With adjustment: all items	% *not* reaching basic LLECE benchmark
Argentina	0	13	2	0	11	9	4
Brazil	0	10	3	0	10	8	4
Chile	0	2	0	0	2	1	1
Colombia	0	8	1	0	10	8	3
Costa Rica	0	3	1	0	2	2	1
Ecuador	0	11	2	0	10	7	3
Guatemala	1	21	4	1	19	17	5
Honduras	0	12	3	0	12	9	6
Mexico	0	8	2	0	6	4	2
Nicaragua	1	21	6	0	20	17	7
Panama	1	16	5	0	16	13	8
Paraguay	2	26	7	1	26	23	11
Peru	0	12	2	0	13	11	4
Dominican Rep.	3	37	12	2	38	35	20
Uruguay	0	10	2	0	9	7	3

Note: Pupil weights not used in this analysis.

Because it was not possible to distinguish incorrect responses from null responses in the publicly available LLECE data, equation (1) could not be applied. However, if the null response patterns in the LLECE data were similar to those in TIMSS and SACMEQ – there is no reason to believe they would not be – then even if equation (1) were applied, this would not make a substantial difference to the findings. The adjusted '% zero' values in Table 7 would drop only slightly, in no cases by more than one or two percentage points. The floor effects problem would still exist.

As in the case of SACMEQ, we must ask whether different proficiency thresholds have been put forward which would change the picture. Altinok (2017: 38) did not use LLECE 2013 data for Grade 3, but did estimate proportions proficient statistics at the at the lower primary level for five Latin American countries, using their performance according to PIRLS and TIMSS.

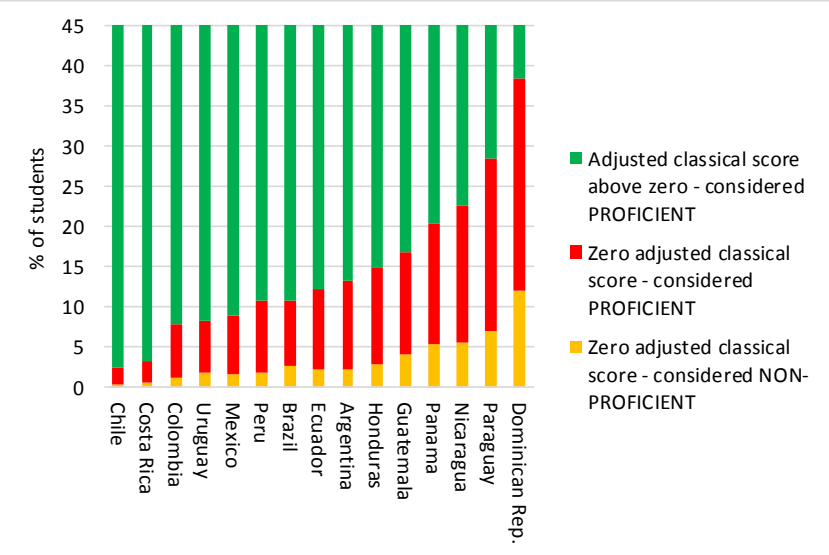
[13] This was verified by examining the spreadsheet accompanying Altinok (2017).

[14] In the LLECE data for Grade 3 mathematics, the last two items described in UNESCO (2016: 139) were missing. For the current analysis, these two items were excluded from *both* the numerator and denominator of the classical scores. The missing items would not affect the analysis in any noteworthy way.

Those statistics are largely in line with what is shown in Table 7, meaning they do not change the picture.

The following graph sums up the problem, using Grade 3 mathematics figures from Table 7. For the purposes of understanding floor effects, students can be divided into the three groups shown in the graph. What is obviously problematic is the middle group: students with an adjusted classical score of zero who are considered proficient. These students may have higher IRT scores than those in the bottom group (those considered non-proficient), but after controlling for random guessing, we essentially do not have enough information on students from both groups to say much about what that can and cannot do. This does not mean that the proportion proficient statistics are not meaningful. If one ranks countries by the published statistics, one obtains almost the same results as if one ranks countries by the percentage of students with an adjusted score of zero. The key point is that the LLECE tests do not include enough easy items to produce meaningful information about the most marginalised students.

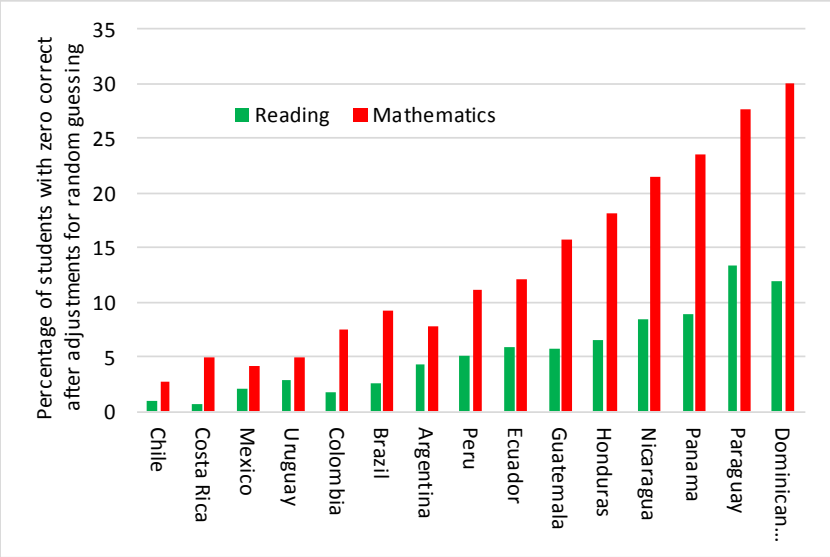
Figure 19: Floor effects in 2013 LLECE Grade 3 reading



Note: Pupil weights used.

Turning to Grade 6, in LLECE, as in SACMEQ, the mathematics data produce higher 'under the floor' values than the reading data – see Figure 20 below. As in SACMEQ, similar IRT scores across the two subjects mask the fact that the mathematics test was considerably more difficult than the reading test. In Dominican Republic, the unadjusted classical score average in mathematics was 26%, against 44% in reading.

Figure 20: Floor effects in 2013 LLECE Grade 6



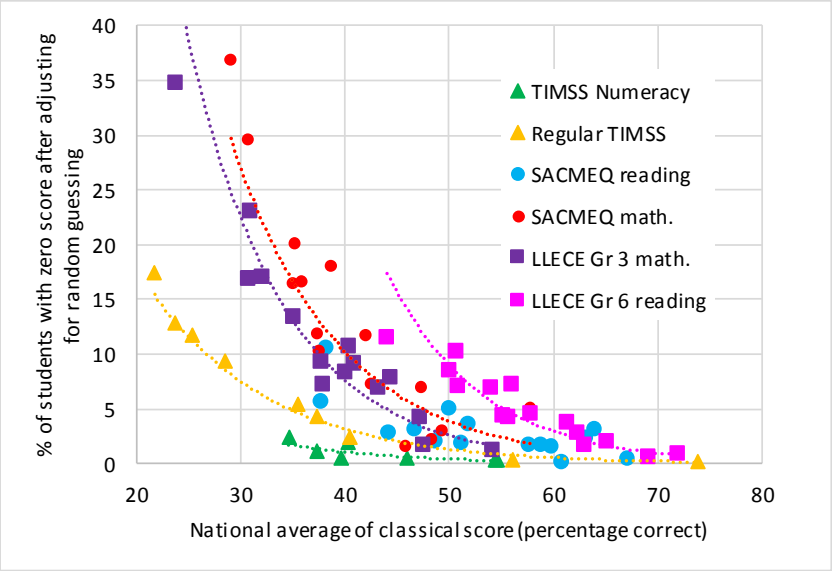
Note: Pupil weights used.

In LLECE Grade 6, as in LLECE Grade 3, both subjects discussed above display 'under the floor' figures which exceed LLECE's own proportions proficient statistics. Floor effects are thus a problem in all the four largely multiple choice LLECE tests analysed here.

The proportions proficient in the Altinok (2017) dataset make no difference to the Grade 6 conclusions. The 'basic' thresholds used by Altinok, which are based on SACMEQ's thresholds, are slightly higher (more demanding) than what one finds in LLECE in the case of mathematics, and slightly lower (less demanding) in the case of reading.

Figure 21 below provides a sense of what the classical means in tests should be if substantial floor effects are to be avoided. National figures that draw from the preceding analysis of the TIMSS, SACMEQ and LLECE data are used. Clearly, relatively high classical scores must be achieved if there are no or very few constructed response questions. The extensive use of constructed response items in TIMSS is what allows for much lower classical means in TIMSS. In SACMEQ and LLECE, either more constructed response items should be introduced, or easier multiple choice items should be introduced that would ensure that no country scored below approximately 50% correct. That classical score would in general limit students with an adjusted score of zero to 5% or less, which is arguably tolerable. The graph suggests that reducing floor effects even further would be difficult without introducing constructed response items. It would be understandable if programmes such as SACMEQ wish to avoid this given the additional costs and complexity, but also risks in relation to the comparability of the scoring process, of having constructed response items. What this discussion points to, is that while the very large floor effects seen in some countries can be eliminated, completely eliminating these effects would be costly and would involve substantially different psychometric approaches.

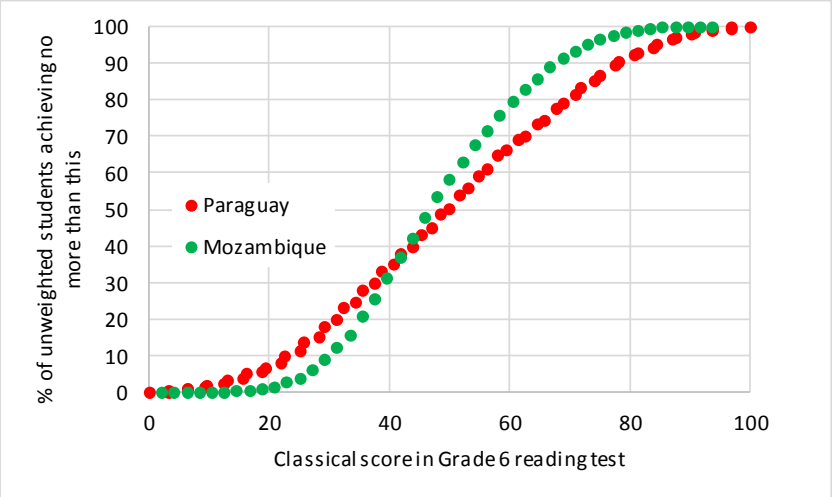
Figure 21: Random guessing effects in TIMSS, SACMEQ and LLECE



Note: Pupil weights not used. Trends are exponential.

Of course, it is not just a more suitable classical score mean which must be aimed for. The distribution of scores impacts of floor effects too. In Figure 21, LLECE Grade 6 reading emerges as particularly prone to floor effects. The reason why can be seen in the following graph, which compares SACMEQ and LLECE reading scores across Paraguay and Mozambique. Both countries display similar classical means – 51% in Paraguay, against 49% in Mozambique. Yet when classical scores are adjusted to control for random guessing, 10% of students in Paraguay emerge with a score of zero, against 2% in Mozambique. It is clear from Figure 22 that despite having a slightly lower classical mean, Mozambique succeeds in having fewer students with very low scores, which reduces the floor effect.

Figure 22: Paraguay in LLECE compared to Mozambique in SACMEQ



Note: Pupil weights not used.

To complete this section, we ask the question of how easy the LLECE tests would need to be to bring floor effects down to a minimal level. What would happen if, say, Grade 6 mathematics students in Dominican Republic wrote the SACMEQ tests? This can be estimated, with a fair degree of accuracy. Altinok's (2017) proportion proficient statistics put Dominican Republic roughly on a par with Zimbabwe, with respect to mathematics. Dominican Republic is the

weakest performer in LLECE, while Zimbabwe is an above average performer in SACMEQ. This serves as a reminder of an important fact: Latin America performs better than the Southern and East Africa region educationally. Dominican Republic's average classical score in Grade 6 mathematics in LLECE was 25%, while Zimbabwe achieved 42% in the SACMEQ test. Even with this easier test, Dominican Republic would have displayed substantial floor effects (see Zimbabwe's 11% 'under the floor' in Figure 17), suggesting that ideally a test where the average item was even easier than in SACMEQ would be required to bring about an adequate assessment of weaker students in Dominican Republic.

6.4 The extent of the 'non-assessed'

All of the preceding analysis has considered floor effects using only data on students who participate in the assessment. What has not been considered, are children who for some reason were not assessed. As will be explained, these children are largely the most disadvantaged in society. This means that the statistics illustrating the extent of students 'under the floor', for instance Figure 17 for SACMEQ, are under-estimates insofar as they do not consider disadvantaged 'non-assessed' children. The current section examines the extent of this, with a special focus on those countries which, in the preceding analysis, emerged as having the largest floor effects.

Three categories of the 'non-assessed' can be identified. Firstly, there are children who are excluded through the rules of the assessment programme. For instance, virtually all assessment programmes exclude children in special needs institutions as it is felt that the assessments are not designed for many of these children. Moreover, some programmes also exclude children in small and remote schools in order to reduce costs. Secondly, there are children who are excluded because they are absent on the day of the test. Thirdly, there are children who are not attending any education institution. The three categories are discussed below.

TIMSS rules permit exclusions of up to 5% of students and 2% of schools due to factors such as disability, the unusualness of the student's language and the smallness of the school. Overall exclusions in TIMSS Grade 4 can be high, sometimes exceeding the TIMSS limits, for instance 10% of students in Singapore and 7% in the United States. The high level of exclusion in Singapore is due mainly to the exclusion of relatively elite international schools which do not follow the national curriculum. Among the TIMSS developing countries discussed in the current report, exclusions ranged from 0.2% in Indonesia to 5.6% in Bahrain. Figures were 2%, 4% and 4% for South Africa, Iran and Chile (Martin *et* al, 2016: 3.7, 5.8, 5.139).

As mentioned earlier, technical documentation for SACMEQ 2007 is scarce. However, the SACMEQ 2000 documentation, which is better, is likely to reflect the situation in more recent waves of the programme. In 2000, exclusions within 'mainstream registered' schools, largely resulting from the exclusion of small schools, amounted to between a reported 0%, for Mozambique, and 4%, for many countries including South Africa, Kenya and Malawi. The figure for Zambia, which displayed the largest floor effects in Figure 17, was 3% (Chimombo *et al*, 2005: 32).

No international dataset appears to exist describing the extent of enrolment in special schools, outside mainstream schools. Country-specific reports point, for instance, to 0.9% of all students being in special schools in South Africa, and 0.1% in Zambia. Zambia's figure can probably be considered typical for Africa. These students would be among the non-assessed in SACMEQ (South Africa, 2016; Serpell and Jere-Folotiya, 2011: 217; Hegarty, 1994: 14).

LLECE documentation indicates that the exclusion of special schools was applied. Moreover, in each country, the smallest 2% of schools were excluded to reduce costs. This would obviously translate to much fewer than 2% of students. Importantly, LLECE 2013 included some students who had to be tested in a language other than Spanish or Portuguese, an

improvement on LLECE 2006 (known as 'SERCE'), where only these two languages were used (UNESCO, 2016: 201-3). Statistics from a couple of Latin American countries appear to confirm that less developed countries tend to have lower proportions of students in special schools, probably because many disabled children are not in any institution at all. In Bolivia, special schools have accounted for 0.3% of school enrolments, against 1.9% in Chile (Peredo, 2012; Chile, 2018: 31).

The second category of the non-assessed is students who are absent on the day of the test. Clearly, the higher the levels of absenteeism, the less reliably results will reflect the proficiency of the student population. However, biases in the results, and in statistics on floor effects, become particularly serious if on average students who are absent perform differently in school to those who are not absent. Table 8 draws from student responses in SACMEQ to the question of how many school days were missed during the last full month. In a few countries, such as Zambia, the figures are high: the mean of 2.5 days translates into an average absenteeism rate of around 12%. What is critical, though, is Figure 23, which confirms that students with a lower level of performance in tests tend to display higher levels of absenteeism. This is absenteeism in general, not absenteeism on the day of the test, which may be unusually high or low, depending on how schools react to knowing the test will occur. However, there appears to be no data on absenteeism on the day of the test, so general absenteeism patterns must be used for the analysis. The patterns seen in Figure 23 suggest that absenteeism will tend to bias results upwards, and make floor effects appear slightly lower than they in fact are.

Table 8: Student absenteeism for Grade 6 in SACMEQ 2007

	Average days absent in a month per student	Median	75th percentile	90th percentile
Botswana	0.4	0	0	1
Kenya	1.3	0	2	4
Lesotho	1.5	0	2	5
Malawi	1.7	1	3	5
Mauritius	1.8	1	2	4
Mozambique	1.1	0	2	3
Namibia	1.0	0	1	3
Seychelles	1.7	1	3	5
South Africa	1.0	0	1	3
Swaziland	0.4	0	0	1
Tanzania	2.1	1	3	6
Uganda	2.4	2	4	6
Zambia	2.5	2	3	6
Zanzibar	1.8	1	2	5
Zimbabwe	1.7	0	2	5

Figure 23: Student absenteeism and performance in SACMEQ 2007

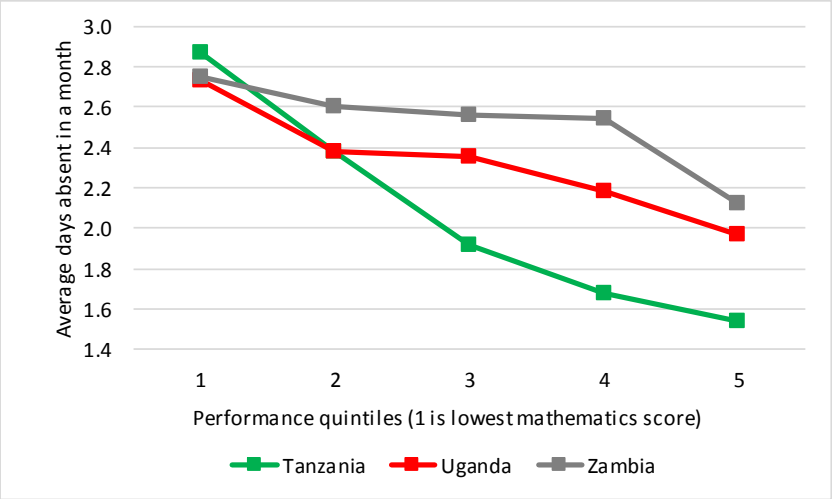
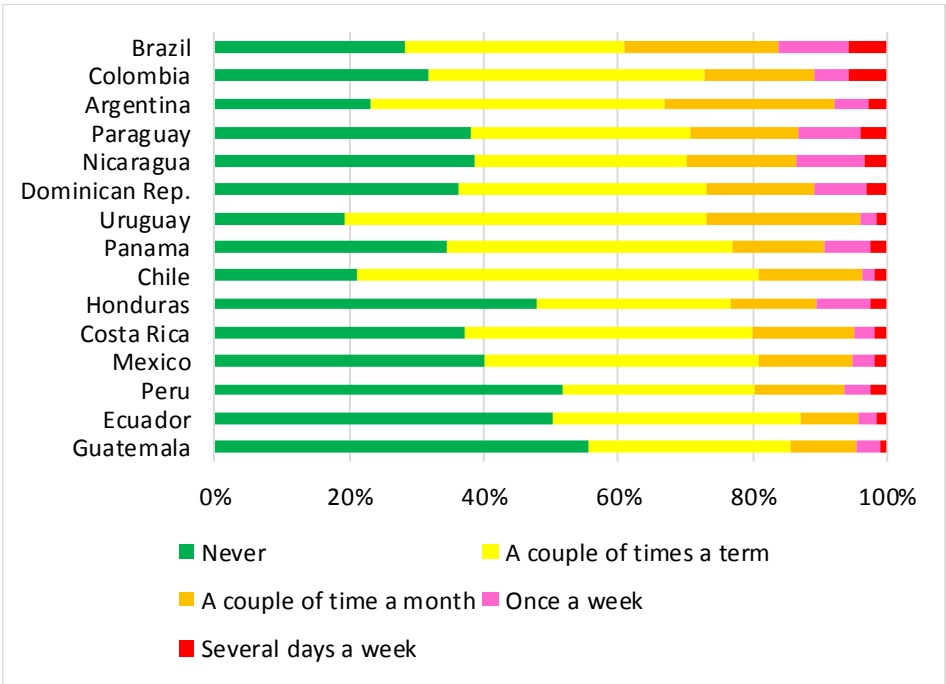


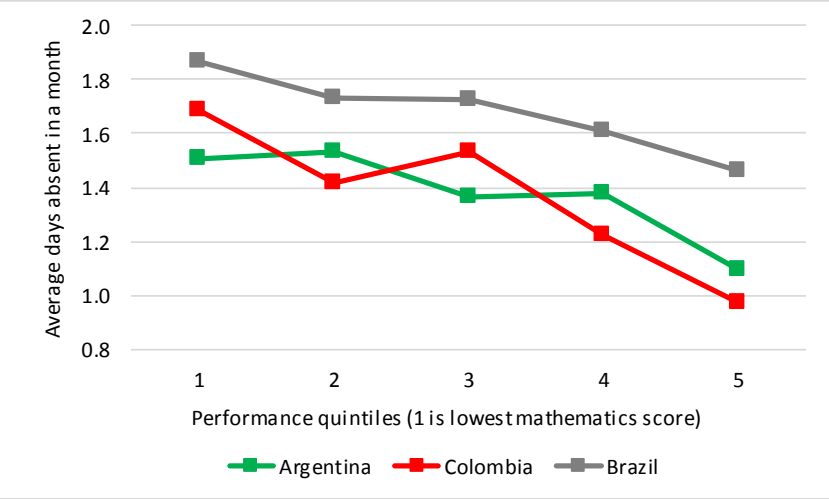
Figure 24 illustrates Grade 6 student responses in LLECE to a question on how often they missed school over the last six months. Figure 25 confirms what one would expect, namely that in the Latin America region too, higher levels of absenteeism are associated with lower academic performance. As in Southern and East Africa, absenteeism would result in an under-estimation of floor effects.

Figure 24: Schools days missed by Grade 6 students in LLECE



Note: Countries are sorted according to an estimate of days absent per month, with the five categories corresponding to 0, 1, 2, 4 and 8 days. Brazil displayed the highest value, at 1.7 days a month.

Figure 25: Student absenteeism and performance in LLECE 2013



Note: Average days in a month is estimated using the assumptions appearing in the note to the previous graph.

The third category of the 'non-assessed' is children who are not enrolled in school. UIS.Stat's variable 'Rate of out-of-school children of primary school age', where the data source is household data, easily the most reliable data for this, points to rates ranging from 1% to 5% for LLECE countries, and 2% to 13% for SACMEQ countries[15]. The three worst countries, by far, are Kenya (12% in 2014) and Uganda and Zambia (both 13%, for 2016 and 2013 respectively). Notably, there was no recent statistic for Mozambique.

Clearly, the statistics lined up in this section compound the marginalisation of students in certain countries, beyond what was seen in previous sections. To illustrate a rather problematic country, in Zambia we can conclude that around 28% of children who should be in Grade 6, were not tested in SACMEQ. This is 13% out-of-school, plus 3% in schools too small to be included in SACMEQ, plus 12% who are enrolled but were absent from school. This last 12% is only partly worrisome, insofar as absent learners were worse performers on average than students who attended school on the day of the test. From a policy and monitoring angle, the out-of-school and even the small school students should be more concerning. These two categories come to 16%. This extent of the 'non-assessed' can be compared to the around 31% of *children* (not students) who are *under*-assessed in the sense that they were tested but ended up 'under the floor' in the 2007 SACMEQ mathematics test. This leaves one with just 53% of children who were properly assessed, in the sense that they were given a meaningful test score. This is a rather striking way of acknowledging how limited the monitoring of academic results is in some countries.

The situation is clearly not as bad in many other developing countries, such as Swaziland (now called Eswatini) or Chile. In fact, as countries develop, increasingly their largest 'non-assessed' problem becomes children enrolled in special schools.

To conclude, this section has provided a more comprehensive view who 'falls through the cracks' in existing monitoring systems. It has also showed that beyond the test programme data, information is often scarce. And it has pointed out how complex the full picture is, encompassing various categories of children.

[15] UIS.Stat consulted in May 2019. The year range was 2012 to 2018, with many cells being blank.

7 Conclusion for policymakers

This paper has presented an analysis of floor effects in international testing programmes with the aim of informing the way forward for monitoring learning outcomes, and in particular the monitoring of SDG Indicator 4.1.1, which tracks the proportion of students at various stages of the schooling system reaching minimum proficiency levels. The methods put forward in the paper for gauging the extent of floor effects in existing assessments could almost certainly be improved upon, yet they represent an important step towards understanding an under-analysed yet important phenomenon. Without a proper understanding of floor effects, designers of assessments may produce systems which are not fully fit for purpose, important faults in existing assessments may continue without being corrected, and analysts may interpret results incorrectly.

Though programmes such as SACMEQ and LLECE focus specifically on developing countries, and were designed to cater in better ways for their weaker educational performance, these programmes often display large floor effects. Put differently, these programmes are in the case of at least some subjects and grades worse than many would believe at assessing weaker learners well. To illustrate, with respect to Grade 6 mathematics, in countries such as Zambia (in SACMEQ) and Dominican Republic (in LLECE), around a third of tested students effectively score zero after one has controlled for random guessing in multiple choice questions. Yet the proportion of students performing under minimum proficiency levels set by these programmes is well below a third. These proportions are based on scores where there has been no adjustment for random guessing. This in effect means that children about which we know very little, in terms of their abilities, are being put into categories such as proficient and non-proficient. This is clearly not a desirable situation.

The uncertainty brought about by floor effects is not large enough to substantially change the rankings of countries in programmes such as SACMEQ and LLECE, regardless of the ranking measure used. However, what is concerning is that this uncertainty substantially reduces our ability to monitor a country's progress over time. Even relatively good progress involves small changes in results over time and thus requires monitoring tools without large margins of error. To illustrate, floor effects of the kind seen in several developing country results make it virtually impossible to gauge whether even impressive improvements have occurred in any period of less than ten years (section 4). And this is before taking into account margins of error associated with the sampling nature of the data.

Attainment of the SDG goals on learning proficiency requires effective monitoring. Without such monitoring, important improvements could be occurring 'under the radar', with existing assessment systems unable to detect this. This can contribute towards policy instability, something which schooling systems are prone to. Without evidence on improvements, it is easier for lobbyists of new policies to succeed, and hence for policy instability to occur, which in turn could slow down improvements. Put differently, a vicious cycle of inadequate knowledge and continuous policy change could be perpetuated.

Moreover, weaknesses in the ability of assessment programmes to gauge what is happening can be exploited by stakeholders opposed to standardised testing for some reason. On a practical level, the problem with floor effects is that having very little information on the most disadvantaged children in society means that the exact learning challenges they face will not be properly understood. This in turn makes it more difficult to come up with better interventions, and in particular teaching methods, which will address these challenges.

What is to be done in programmes such as SACMEQ or LLECE to reduce floor effects? This question is almost certainly relevant to many other international programmes focussing on developing countries, such as PASEC, and to many national programmes. At least in theory, one solution is to introduce more constructed response questions. A large reason why a

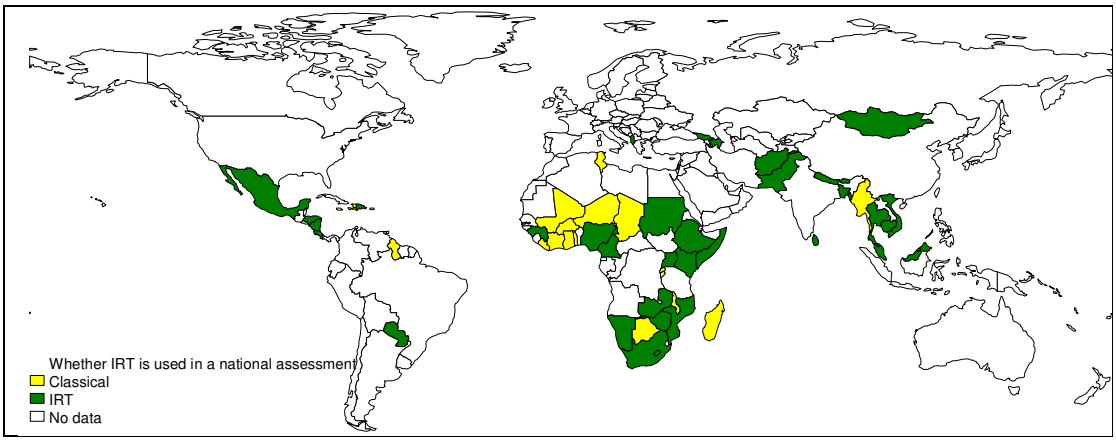
programme such as TIMSS is better at dealing with floors is that around half of the questions are constructed response questions, or questions where a marker must look at a constructed response and no random guessing is possible. SACMEQ has no constructed response items, and LLECE just a few. However, introducing such items, or expanding their presence, is costly, complex and risky. A separate writing test in LLECE demonstrates this. This test, which has no multiple choice questions and involves writing a short text, appears to be weak at gauging the abilities of students. This is probably because of difficulties in standardising marking sufficiently. Such standardisation is not impossible, but it requires advanced and complex procedures.

The introduction of TIMSS Numeracy in 2015 in Grade 4 provides a useful demonstration of how adding more easy multiple choice items can reduce floor effects. The proportion of children 'under the floor' in TIMSS Numeracy is less than half of that found in regular TIMSS, counting similar children and just the multiple choice part of TIMSS. TIMSS Numeracy was not able to eliminate entirely the floor effects in this part of TIMSS, and eliminating floor effects in a test consisting entirely of multiple choice items would be difficult. However, there is clearly scope for reducing floor effects in programmes such as SACMEQ and LLECE without introducing constructed response items.

Much of this paper has focussed on the classical scores behind the item response theory (IRT) scores which are typically the focus when results are presented from assessment programmes which use IRT. Analysts should be encouraged to explore what lies behind IRT, in part because this helps to 'demystify' IRT. Classical scores are what teachers understand. Yet they need to understand IRT to a greater degree, in particular as standardised assessments using IRT become more widespread. This is one way of contributing to the professionalisation of teachers and ensuring that teachers can identify with national and global efforts to improve and monitor learning outcomes. IRT helps to provide richer information on what students can and cannot do. But it can also mask floor effects and create the illusion that these effects are not a problem. Put differently, IRT scores should not be used as a way of concealing the fact that certain tests are more difficult than they should be. For IRT to be used correctly, and for its full potential to be realised, knowledge about it should become more widespread, especially in developing countries. This paper has confirmed that IRT is easily used incorrectly, and that not having clear technical documentation on IRT processes which were followed is a part of this problem.

To end the paper, the following map draws from a UIS online dataset specifying the type of scoring used in national assessments in developing countries. Of the 53 countries for which data were available, 34 had some assessment where IRT was reportedly employed, while 19 countries used only classical scores. For those familiar with assessments in a few countries it should be clear than even in the countries coloured green, IRT is often used in a limited or even inappropriate manner, or was used at some point in the past, and then discontinued. The need to develop the capacity for wider use of IRT is clearly a reality.

Figure 26: Developing countries which have used IRT in a national programme



Source: UIS 'Database of learning assessments', at http://uis.unesco.org/en/uis-learning-outcomes. Accessed January 2019.

References

Altinok, N. (2017). *Mind the gap: Proposal for a standardised measure for SDG 4 – Education 2030 agenda.* Montreal: UIS.

Altinok, N, Angrist, N & Patrinos, H.A. (2018). *Global dataset on education quality* (1965-2015). Washington: World Bank.

Australian Curriculum, Assessment and Reporting Authority (2017). *National Assessment Program – Literacy and Numeracy 2017: Technical report*. Sydney.

Burton, R.F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education,* 26(1): 41-50.

Carnoy, M., Khavenson, T., Fonseca, I. & Costa, L. (2015). Is Brazilian education improving? Evidence from Pisa and Saeb. *Cadernos de Pesquisa*, 45(157).

Catts, H.W., Petscher, Y., Schatschneider, C. *et al* (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities*, 42(2): 163-176.

Chile (2018). *Estadísticas de la educación 2017.* Santiago: Ministerio de Educación.

Clarke, M. (2012). *What matters most for student assessment systems: A framework paper*. Washington: World Bank.

Chimombo, J., Kunje, D., Chimuzu, T. & Mchikoma, C. (2005). *The SACMEQ II project in Malawi: A study of the conditions of schooling and the quality of education*. Gaborone: SACMEQ.

Crouch, L. & Gustafsson, M. (2018). *Worldwide inequality and poverty in cognitive results: Cross-sectional evidence and time-based trends*. Oxford: Research on Improving Systems of Education.

Gustafsson, M. (2014). *Education and country growth models*. Stellenbosch: University of Stellenbosch.

Hanushek, E.A. & Woessmann, L. (2009). *Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation*. Washington: National Bureau of Economic Research.

Hastedt, D. & Desa, D. (2015). Linking errors between two populations and tests: A case study in international surveys in education. *Practical Assessment, Research & Evaluation,* 20(14): 1-12.

Hegarty, S. (1994). *Educating children and young people with disabilities*. Paris: UNESCO. Available from: <http://inee-assets.s3.amazonaws.com/resources/Educating_with_disabilities.pdf> [Accessed May 2016].
Izard, J. (2005). *Trial testing and item analysis in test construction*. Paris: IIEP.
Jerrim, J. (2013). The reliability of trends over time in international education test scores: Is the performance of England's secondary school pupils really in relative decline? *Journal of Social Policy,* 42(2): 259-279.
Kellaghan, T., Greaney, V. & Murray, T.S. (2008). *Using the results of a national assessment of educational achievement*. Washington: World Bank.
Makuwa, D.K. (2010). Mixed results in achievement. *IIEP Newsletter,* XXVIII(3).
Martin, M.O., Mullis, I.V.S., Hooper, M. (eds.) (2016). *Methods and procedures in TIMSS 2015.* Chestnut Hill: IEA.
Martin, M.O., Mullis, I.V.S., Hooper, M. (eds.) (2017). *Methods and procedures in PIRLS 2016.* Chestnut Hill: IEA.
Moloi, M.Q. & Chetty, M. (2011). *Trends in achievement levels of Grade 6 pupils in South Africa*. Paris: IIEP.
Mullis, I.V.S., Martin, M.O., Foy, P. & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill: IEA.
Mullis, I.V.S., Martin, M.O., Foy, P. & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Chestnut Hill: IEA.
Mullis, I.V.S., Martin, M.O., Loveless, T. (2016). *20 years of TIMSS: International trends in mathematics and science achievement, curriculum and instruction*. Chestnut Hill: IEA.
Peredo, R. (2012) *La situación de la educación especial a través de datos y indicadores educativos.* La Paz: Universidad Católica Boliviana.
Pritchett, L. & Beatty, A. (2012). *The negative consequences of overambitious curricula in developing countries*. Washington: Center for Global Development.
Resch, A. & Isenberg, E. (2014). *How do test scores at the floor and ceiling affect value-added estimates*. Princeton: Mathematica Policy Research.
Serpell, R. & Jere-Folotiya, J. (2011). Basic education for children with special needs in Zambia: Progress and challenges in the translation of policy into practice. *Psychology and Developing Societies,* 23(2): 211-245.
South Africa (2016). *Education statistics in South Africa in 2014.* Pretoria: Department of Basic Education.
UNESCO (2014). *Comparación de resultados del Segundo y Tercer Estudio Regional Comparativo y Explicativo: SERCE y TERCE 2006-2013*. Santiago.
UNESCO (2016). *Reporte técnico: Tercer Estudio Regional Comparativo y Explicativo*. Santiago.
UNESCO (2018). *SDG 4 data digest 2018: Data to nurture learning.* Paris.
UIS (2017a). *Principles of Good Practice in Learning Assessment*. Montreal.
UIS (2017b). *Constructing UIS proficiency scales and linking to assessments to support SDG Indicator 4.1.1 reporting*. Montreal.
UIS (2018). *Costs and benefits of different approaches to measuring the learning proficiency of students (SDG Indicator 4.1.1)*. Montreal.
World Bank (2018). *World Development Report 2018: Learning to realize education's promise*. Washington.