CENTER FOR GLOBAL DEVELOPMENT

Psychometric Quality of Measures of Learning Outcomes in Low- and Middle-Income Countries

Masha Bertling, Abhijeet Singh, and Karthik Muralidharan

Abstract

We investigate the properties of measures of learning outcomes, as these are the tools commonly used to monitor the progress toward identifying the most effective interventions. We review test properties across 158 studies and conduct item-level psychometric analysis of a subset of these studies to show that current tests vary widely in scope, content, administration, and analysis. Researchers rarely provide details about the properties of their test scores. Only 4% of studies we review provide reliability estimates of their tests, and 10% archive item-level replication data to evaluate test quality post hoc. The interpretation of any estimates is necessarily sensitive to the measurement of the core variables, even where treatments are randomly assigned. Since estimates of treatment effects are often expressed in standard deviation units, measurement error can bias treatment effects toward zero. Content analysis of question wordings reveals substantial variation in content coverage of the skills tested, even when students of similar grades are being tested in similar subjects. The findings indicate that comparisons of treatment effects must consider degrees of measurement error that are often unavailable and the content breadth of the tests to contextualize why effects may differ on substantively different outcome variables.

WORKING PAPER 638 • MARCH 2023

Psychometric Quality of Measures of Learning Outcomes in Low- and Middle-Income Countries

Masha Bertling
Harvard Graduate School of Education

Abhijeet Singh
Non-resident fellow, Center for Global Development

Karthik Muralidharan
University of California, San Diego

We are grateful to Andrew Ho, Justin Sandefur, Susannah Hares and Lee Crawfurd for comments on an earlier version. We acknowledge support from the Center for Global Development education research consortium, funded by the Bill and Melinda Gates Foundation, and the Mastercard Foundation. We are also grateful to authors of the studies in the systematic review for sharing microdata. Andrew Avitable provided painstaking research assistance. All authors participated in conceptualization of the study; MB led the analysis of the microdata in the paper; MB and AS co-wrote the draft.

Masha Bertling, Abhijeet Singh, and Karthik Muralidharan. 2023. "Psychometric Quality of Measures of Learning Outcomes in Low- and Middle-Income Countries." CGD Working Paper 638. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/psychometric-quality-measures-learning-outcomes-low-and-middle-income-countries

CENTER FOR GLOBAL DEVELOPMENT
2055 L Street, NW Fifth Floor
Washington, DC 20036

1 Abbey Gardens
Great College Street
London
SW1P 3SE

www.cgdev.org

Contents

1. Introduction **1**

2. The current state of assessments **4**

2.1 Analytical strategy 4

2.1.1 Systematic literature review 4

2.1.2 Item-level analysis 5

3. Results **7**

3.1 Systematic literature review 7

3.1.1 Researcher-designed assessments 8

3.2 Item-level analysis 9

3.2.1 Content analysis 9

3.2.2 Psychometric analysis 11

4. Recommendations for practice **13**

4.1 Case study: Muralidharan, Singh and Ganimian (2019) 13

4.2 Recommendations 15

5. Conclusions **16**

References **18**

Figures and tables **20**

Appendix **28**

Codebook 28

List of Figures

1. Country representation in systematic review 20
2. Subdomains covered in test forms in literacy and math 21
3. Distribution of reliability estimates in math and literacy 22
4. Item difficulty and test information across subdomains 22
5. Presence of floor and ceiling effects 23
6. Sample math item from a test form with a 96.5% of the floor effect 23

List of Tables

1. The systematic review sample (top panel) and researcher-designed assessments' characteristics (bottom panel) 24
2. Quality of a test form and various item characteristics: descriptive statistics of the variables 26
3. Estimated effects of item characteristics on test quality, controlling for grades 27

1. Introduction

Developing countries have made rapid gains in access to schooling, but levels of achievement remain very low (World Bank, 2017). Remedying this "learning crisis" has emerged as a focus of research in the economics of education (World Bank, 2017, p.4). In low- and middle-income countries (LMICs) over the past 15 years, there have been several hundred policy evaluations that have aimed to identify interventions that improve student achievement (Connolly *et al.*, 2018; Muralidharan, 2017). Student achievement, measured through test scores on standardized assessments, is a key outcome measure of this large (and growing) literature.

In high-income countries, test scores nearly always come from secondary data sources such as the National Longitudinal Surveys (NLSs), the Programme for International Student Assessment (PISA) assessments, or administrative data collected by schooling systems. These assessments were designed and administered independently, frequently involving large teams of psychometricians and testing experts. While economists using these data must make meaningful choices about analyzing the data (see Jacob and Rothstein (2016)), they do not typically control what the tests assess, how they are administered, or how they are scored. In contrast, development economists often field their own assessments, primarily due to the lack of tests that would provide information at the lower end of the achievement distribution. While this provides substantial opportunities to tailor assessments to the relevant population and research question, it also raises the risk of psychometrically unsound designs, which reduce the precision or meaning of treatment effects within and across studies. Unlike many aspects of survey design (see, e.g., Grosh and Glewwe (2000)), no authors in our sample directly cited standards for education testing provided by American Educational Research Association *et al.* (2018).

This paper reviews current practices in development economics relating to the design of educational assessments and suggests practices to improve these assessments. We incorporate factors specific to the analytical goals of these studies and the economics literature, as well as constraints posed by researcher time, survey length and complexity, and the low-and-dispersed levels of student achievement in many settings. We attempt this in three steps.

First, we conducted a review of 158 studies and classified them on a range of 95 characteristics, including psychometric and content-related assessment characteristics and broader sample and study characteristics. We use a coding scheme that reflects criteria in modern psychometric standards (American Educational Research Association *et al.*, 2014). We selected studies by whether they were featured in prominent systematic reviews (Kremer *et al.*, 2013; Glewwe and Muralidharan, 2016), published in one of five prominent economics journals, or conducted by a J-PAL affiliate in the past ten years. While we do not claim to have comprehensive coverage of all international education policy evaluations, we believe our sample is adequate to characterize current practices in the economic literature.

Next, we collected and analyzed item-level data from a sample of 40 studies alongside the administered test forms. We coded all test items (N = 5944) in the studies for which we received question wordings to classify the specific competency that each item measured. Then, we conducted a psychometric analysis of each test booklet using both classical and Item Response Theory (IRT) models. The aim of this exercise was three-fold: to subject test booklets from different studies and settings to a consistent set of data quality checks; to assess the diversity in what researchers test, even within a single broad domain (such as "Mathematics"); and, finally, to assess whether it is feasible ex-post to link test scores across studies and put student achievement on the same scale, hence, allowing for absolute comparisons of effects across studies.

Our first result is that the studies we reviewed rarely contain information about the test design and administration, including what specific skills were tested, how tests were administered, how they were scored, and how the test performed in practice in the study sample. In the absence of such information, the published studies themselves are an insufficient guide to assess whether results may be compared reliably with previous studies or what treatment effects, typically expressed in internally standardized z-scores, mean in absolute terms. Only 9.5% of studies, even when archiving replication data files for public access, include item-level scores or test instruments that would allow for forming such an assessment independently post-publication.

Second, our review of the question wordings of individual questions reveals substantial variation both in content coverage of the skills tested, as well as the modes of administration (e.g., whether students respond to oral or visual stimuli provided individually by proctors or only to written paper-and-pencil tests), even when students of similar grades are being tested in similar subjects. Combined with other differences in the test instruments and samples, this considerable variation in actual questions administered implies that any attempts to formally link test booklets across studies are unlikely to succeed. Put simply, there is no sufficiently large common bank of items that may be used to link assessments across the literature that would allow researchers to put student achievement on a common metric using the current studies alone.

Third, our analysis of the psychometric properties of the test booklets for the studies for which we received item-level information does provide some reassurance about the actual state of practice in most of the literature. Although they may lack comparability with previous work, most studies have internally coherent test instruments. One standard indicator of score quality is Cronbach's alpha (Cronbach, 1951), which estimates the correlation between reported scores and scores on a replicated test that is comprised of the same number of similar items. We estimate Cronbach's alpha to be about 0.7 across the tests in our sample, on average. While this is lower than the degree of internal consistency seen in assessments in the US (typically with alpha coefficients above 0.90, see Reardon and Ho, 2015) or in leading international assessments like PISA and TIMSS (marginal reliabilities—the IRT equivalent of alpha—above 0.90; PISA 2018 technical report, chapters 9 and 12), it indicates that the tests typically administered by researchers in development economics are likely to be reliable for population-level inference which is the goal of most economics research.

That being said, many test booklets do show low alphas and/or display substantial floor effects (i.e., students answer every question on a test). This is wasteful from the perspective of information gathered about students' proficiency and shows that a wiser test design can improve precision and potentially enable shorter tests, hence, reducing the test-taking burden.

We should also keep in mind that Cronbach's alpha only addresses one type of reliability, item-to-item generalizability. Due to the RCT designs, where we measure outcomes at multiple time points (e.g., midline and endline) or rely on multiple different raters to administer the assessments, other types of reliability might be explored to explicitly account for the design effects. That said, careful piloting and item selection could and should improve the informativeness and reliability of tests used by economists in developing countries.

The main contribution of this paper is to provide the first systematic overview of measurement practices in the economics of education RCTs in LMICs—and to highlight areas for improvements that may be possible even within existing resource and context-specific constraints. Given the substantial diversity in the content being tested, psychometric properties, and the modes of administration that we find, making comparisons across studies is tricky even when restricted to similar populations. When combined with additional levels of diversity in geographical contexts, education systems, and the age of students being tested, this implies that standardized effect sizes are, without additional context, a poor metric to decide whether program effects are meaningful.

We further contribute to a small but growing literature concerned about constructing comparable measures of learning outcomes in developing countries. There are several measurement challenges researchers need to address. First, we find that many measures have poor quality. Second, many measures have undocumented quality. Third, overlap among items and populations is insufficient to support the ex-post expression of effects on a common scale (see Koch *et al.*, 2015, for an example of successful linking in the US). Technically, it is possible to link measures using external assessments which combine items from multiple tests. However, previous attempts to achieve this, even in a small set of assessments administered at a similar age, suggest that this is likely to be subject to substantial uncertainty in linking (Sandefur, 2018). Unfortunately, our interpretation of the overall literature in this area is that the only robust way to ensure cross-study and sample comparability would be to build in features such as common items (as in, e.g., PISA, TIMSS, or the Young Lives study).

Finally, our review speaks to a broader topic of the importance of public goods in measurement. Measuring global progress on a range of policy targets—such as reductions in the rates of headcount poverty or malnutrition—has required the agreement of global standards on how measures are defined and aggregated. These standards have also been adopted by individual research teams, often in non-representative samples, that aim to make progress on these specific domains. While the quality of learning is now an explicit policy target in the United Nations Sustainable Development Goals, there is no global standard that captures (even for foundational skills acquired in primary schooling) how this is to be measured. Providing a methodologically sound and consistent basis

for such measures—which is also "open source" in allowing researchers to embed the measures or subset of items in their own studies—would enable substantial advances in bringing consistency in measuring policy progress and in the academic literature.

The following section describes the analytical strategy for conducting a systematic review and analysis of item-level data. Section 3 provides results for a systematic literature review of measures of learning outcomes in 158 studies conducted between 2009 and 2020. We next document test characteristics from studies with item-level data. We evaluate both content coverage and psychometric properties of items and test forms. Section 4 concludes.

2. The current state of assessments

2.1 Analytical strategy

2.1.1 Systematic literature review

We characterize current scholarly practice in development economics in two steps. First, we conduct a systematic literature review of measures of learning outcomes published between 2009 and 2020. We relied on three sources as our inclusion criteria:

(i) two prominent reviews of the literature on the economics of education, namely Kremer *et al.* (2013) and Glewwe and Muralidharan (2016), (ii) studies published since 2011 in five of the leading journals which feature development economics articles, namely the American Economic Review; the American Economic Journal: Applied Economics; the American Economic Journal: Economic Policy; the Journal of Development Economics; the Journal of Human Resources, and (iii) impact evaluations conducted by a J-PAL affiliate in the past ten years. We identified and fully coded 158 studies that used academic performance as the key outcome. Studies that focused solely on behavioral academic outcomes, such as dropout rates or attendance, were not included in our final analytic sample. While this is not an exhaustive list of studies conducted in LMIC, we believe this is a population of recent high-quality educational impact evaluations in the development economics literature.

We classified all studies on a range of 95 characteristics, which captured psychometric and content-related assessment characteristics and broader sample characteristics (see the complete list in Appendix A). We developed this coding scheme to reflect modern psychometric standards (American Educational Research Association *et al.*, 2014), which we use as guiding principles or a framework for evaluating the quality of assessments in our sample.

2.1.2 Item-level analysis

The goal of an educational intervention is to detect a treatment effect and communicate its size. When a scale of an outcome measure is well known, such as inches or pounds in physical measurement, or the US's SAT scale score units or NAEP score units, then a simple difference in treatment and control means can often suffice. Absent a well-known scale, treatment effect sizes are translated to standard deviation units. None of the published studies included in our systematic review reported effect sizes on a known scale nor released sufficient information (e.g., common items or common students across studies; Kolen and Brennan (2014) to determine whether the comparability was possible. Therefore, we conducted a search for publicly released item-level data and question wordings as the second step. We relied on online databases such as Harvard Dataverse and openICPSR, and journal websites that publish supplementary materials, for example, American Economic Review. In 19% of cases, public datasets included item-level data and/or question wordings in the language of administration, and the rest reported a version of a normalized aggregated outcome variable. We reached out to the first authors of all studies that used researcher-designed tests, inquiring about the possibility of obtaining raw item-level data and question wordings. In total, we have received and analyzed 40 datasets with 9,000 unique items and over 5,000 question wordings, capturing both quantitative and qualitative information pertaining to individual questions and test forms.

Content analysis. We conducted an in-depth content analysis of the questions to evaluate the degree of alignment of constructs across scales. For each question, we recorded information about the domain, subdomain, and whether the item has been borrowed from other assessments, such as TIMSS or PISA. For evaluating subdomains, we relied on authors' classifications. When this information was unavailable, we qualitatively coded each question following the most common categories to the best of our ability. Finally, we standardized subdomains across categories provided by researchers and our own categorization. For example, if authors labeled an item as "Knows alphabetical sequence in any form," we relabeled it as "Alphabet" to reflect items from different studies that also asked students to recite letters in the alphabet. Alternatively, if researchers coded an item as "read story," we reclassified it as "Reading Ability," following the same logic.

Psychometric analysis. We evaluated various psychometric properties of both test forms and individual items. We began with assessing reliability, which we calculated using Cronbach's alpha as

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_{y_i}^2}{\sigma_x^2}\right) \quad (1)$$

where k refers to the number of items on a test form, $\sigma^2_{y_i}$ is item-specific variance, and σ^2_x is the variance of the observed total score. The alpha coefficient ranges between 0 and 1, with higher values signifying higher internal consistency. High reliability is desirable for assessments since the measurement error undermines the ability to detect the treatment impact by reducing power. When we have an SD unit effect size, measurement error would further bias the effect downward if it is not corrected for reliability, leading to incorrect inference.

Cronbach's alpha addresses only one type of reliability, item-to-item generalizability, and is one of the most popular metrics used to understand the properties and the quality of a measure (Brennan, 2001). Due to the RCT designs, where we measure outcomes at multiple time points (e.g., midline and endline) or rely on multiple different raters to administer the assessments, other types of reliability should be explored to explicitly account for the design effects.

We next relied on IRT to evaluate the properties of individual items and test forms. In contrast to classical methods, where item features are population-dependent, IRT employs multivariate logistic regression methods to obtain item parameter estimates that are theoretically population invariant (van der Linden and Hambleton, 2013). Let the variable Y_{ij} represent the response of examinee j to item i, where $Y_{ij} = 1$ is a correct item response and $Y_{ij} = 0$ is an incorrect response. The item response curve for the two-parameter logistic model (2PL; Birnbaum, 1968) takes the following form

$$\text{logit}[\Pr(Y_{ij} = 1|\theta_j)] = \alpha_i(\theta_j - \beta_i) \quad (2)$$

where θ_j is the individual's proficiency (measured in standard deviation units from a reference population) on a single dimension, α_i is the item discrimination, and β_i is the item difficulty. Assuming conditional independence between responses to the same item across individuals and, conditional on proficiency, between answers to different items by the same person, item parameters are estimated using maximum marginal likelihood estimation techniques.[1]

Discrimination or slope parameter indicates how well a particular item discriminates against different levels of proficiency, θ. Steeper slopes at a particular level of proficiency suggest that it is more discriminative than levels of proficiency with gentler slopes. Theoretically, α parameter estimates range from $-\infty$ to ∞. While negative values are possible, they are considered problematic, suggesting that students with increasing levels of proficiency are less likely to endorse correct responses. This might be due to poor discrimination between proficiency levels or some

1 There is, of course, alternative estimation possible. Bayesian estimation is often used and is more beneficial over maximum likelihood techniques due to its ability to estimate parameters for complex data structures, including hierarchical data or data that violate the basic assumptions of IRT, small samples, as well as parameter estimation in extreme response patterns (van der linden and Hambleton, 2013).

coding error. The difficulty parameter describes how difficult an item is to achieve a 50% chance of endorsing a correct response at a given proficiency level. The higher the value, the more difficult a particular item is. Hence, we focus on the distributions of the average discrimination and difficulty parameters per test form to understand whether the tests collect information for most students, which is desirable for an RCT.

Finally, we ask whether reliability, ceiling effects, and floor effects tend to be higher or lower on average in different grades, subdomains, and administration conditions. Substantial floor (i.e., students answer every question on a test wrong) and ceiling (i.e., students answer every question on a test correct) effects are wasteful from the perspective of information gathered about student's proficiency and show that wiser test design can improve precision and potentially enable shorter tests, hence, reducing test-taking burden. We describe these patterns using the following linear model:

$$Y_{ts} = \alpha + \beta_1 Items_{ts} + \beta_2 Subdomains_{ts} + \beta_3 OAI_{ts} + \beta_4 Grade_{ts} + \epsilon_{ts} \quad (3)$$

where reliability or floor/ceiling effects (Y_{ts}) of test form t in study s is expressed as a function of the number of items ($Items_{ts}$), the number of content subdomains ($Subdomain_{ts}$), the indicator for orally-administered items (OAI_{ts}), the grades the test has been administered in ($Grade_{ts}$), and the error term (ϵ_{ts}).

3. Results

3.1 Systematic literature review

Table 1 (top panel) characterizes the 158 studies in our sample. Most studies are RCTs (70%), conducted across 40 countries, with the largest number coming from South and East Asia (see Figure 1). Only 13% of the studies used nationally representative samples, with most researchers collecting the data in smaller regions of the country or in districts with whom they have historically established relationships. Studies were conducted in both rural and urban areas and primarily in public schools with low-income students. Over half of the studies focused on students in primary grades.

We next classified each measure with respect to its "origin." For over half of the studies, we were able to establish that measures of learning outcomes were developed by research teams solely or in collaboration with teachers and/or external vendors, as opposed to using test scores from the standardized government—or district-level exams—the latter comprised 35% of the studies. Researcher-designed measures are of particular interest for our purposes since researchers have the most flexibility in defining the constructs and developing questions. We identified 73 studies that fit this criterion and examined them in-depth.

3.1.1 Researcher-designed assessments

Whereas almost all researchers reported the primary domain of the test (e.g., mathematics or literacy), we wanted to understand how they defined achievement within it. Hence, we were interested in the subdomains researchers focused on (Table 1, bottom panel). We see that 64% of researchers did not define any construct subdomains for both mathematics and literacy. 14% of those who reported focusing on number properties and simple single-or double-digit manipulations. Literacy tests, in turn, focused on early literacy skills, such as letter or word recognition. This is reflective of most studies conducted in primary grades. Since only a few studies reported information about subdomains, we thought of alternative indicators that can tell us more about content coverage for a domain. We considered two such indicators-alignment to the curriculum and information on whether particular questions were borrowed from various open sources.

First, 61% of researchers indicated that measures of learning they used were reflective of either local or national curricula (Table 1, bottom panel). While curriculum differs across contexts and grades, this suggests that researchers were focusing on broader level skills rather than a narrow subset linked to their intervention. Next, we looked more specifically at whether the questions were developed entirely from scratch or had been borrowed from publicly available sources. Over 43% of research teams indicated that their tests included publicly available questions and were released either as sample questions by OECD assessments (e.g., PISA or TIMSS), governmental exercise books, or national exams. Despite the abundance of this "borrowing" practice, we were not able to definitively identify whether all or a subset of questions were borrowed for each study.

For administration conditions, we were interested to learn whether the tests were administered in schools or at home. This is essential since home visits are costly and potentially less standardized compared to in-school data collection. Almost half of our studies, 49%, administered tests in schools, whereas 23% of research collected achievement data at home. With respect to the modes of administration, 57% of researchers did not specify if students sat for a written test or if a proctor orally administered the test. Those who did, asked students to take either a written paper-and-pencil test or a mix of paper-and-pencil and orally administered tests (22 and 18 percent, respectively). We have also noticed that it was more common to have dual administration modes for studies conducted at home.

Properties of scores and scale construction characteristics. Researchers rarely report psychometric properties of scales. The most common way to evaluate a test property is to look at internal consistency as measured by Cronbach's alpha. Again, Cronbach's alpha estimates a correlation between test scores and scores from a hypothetical test composed of similar items and is only one of the ways to start understanding the properties of a test. Only four percent of studies (three out of 73) that used self-developed tests as outcome measures reported reliability coefficients ranging between 0.65 and 0.90 (see Table 1, bottom right panel). Specifically, Barrera-Osorio and Filmer (2016) reported Cronbach alphas of 0.71 for the math test and 0.65 for the Digit Span test;

Filmer and Schady (2014) reported reliabilities of 0.68 for their math exam and only 0.65 for a cognitive test, the reliability of the vocabulary test was 0.90; Loyalka *et al.* (2019, p. 16) used a math test with a reliability of "approximately 0.80." An additional seven studies used Raven's matrices as one of their outcome measures. These studies reported reliability above .90 or cited the technical report for the instrument. However, in none of these seven studies, the authors provided reliability estimates for all outcomes they used (e.g., math ability).

Since, for most studies, we did not have information about reliability, we thought about proxies that might provide us with information about the quality of the scale. Under classical assumptions, reliability increases with the number of items on a test since it increases the systematic variance in the outcome (Churchill and Peter, 1984; Cronbach, 1951; Jaju and Crask, 1999). In other words, the more items, the higher the reliability. However, inferring differences in reliability solely from differences in test length assumes that inter-item covariances are the same. 29% of studies provided information about the number of items included in either mathematics or literacy assessments at baseline and/or endline, varying from 5 to 98 questions.

The second proxy that we considered relates to the number of items in the test and, hence, reliability—the duration of the assessment. We similarly assume that inter-item covariances are the same and take approximately the same time. We identified 15 studies that provided information about the duration of tests. On average, the tests were 31 minutes long, ranging from 5 to 60 minutes.

3.2 Item-level analysis

3.2.1 Content analysis

We classified 5,994 items for which we had question wordings, relying on broad categories. Within mathematics domain we classified items into the following categories:

(a) "number recognition" category includes items that ask students to name or write down a particular number given visual or hearing clue (e.g., "*Are you able to identify this number (17)?*");
(b) "number concepts" questions ask students to contrast the numbers or identify a missing number to complete a sequence (e.g., "*90; 92; [0]; 96; [ ]*");
(c) "arithmetic" category covers items that ask students to perform, for example, addition, subtraction, multiplication, and/or division operations (e.g., "*Solve* 4/5 + 2/5 * 0.2 + 1.4");
(d) "algebra" category covers operations on functions (e.g., "*Solve the following equation: 4w − 2w = 26: w = "*);
(e) "geometry" category covers items that ask students to identify various shapes and measure its properties (e.g., "*Rama wants to build a fence around his land, which looks as follows. How many meters of fence does he need? Pentagon of perimeter 48m shown. A) 20 Meters B) 24 Meters C) 48 Meters D) 96 Meters*");

(f) "measurement and application" category concerns various units of measurement (e.g., "*The height of the teacher's desk would be about? A) 100 millimeters B) 5 centimeters C) 1 meter D) 2 kilometers*");

(g) "statistics" category covers questions around identifying mean, median, or mode of some variable (e.g., "*Wickets taken by a bowler in 2020 matches played by him are as follows: 1,0,2,3,5,2,4,6,2,0,1,5,3,4,3,3,2,1,2,01,0,2,3,5,2,4,6,2,0,1,5,3,4,3,3,2,1,2,0. Find the mode of the above data.*");

(h) "word problems" category includes questions that might combine concepts from other categories and are presented in a form of a word problem with application to real life (e.g., "*Think carefully about the following problem: A farmer trader has 72 eggs which are to be put in 8 boxes. Each box will contain the same number of eggs. If the price of one egg is 550 rupiah, then the price of one box of eggs is: a. Rp. 4,950; b. Rp. 5,000; c. Rp. 5,250; d. Rp. 5,450*").

We classified literacy items into eight subdomains:

(a) "letter recognition" category covers items that ask students to identify individual letters (e.g., "*Are you able to identify this letter: C?*");

(b) "word recognition" category, similarly, concerns the identification of words rather than single letters (e.g., "*Circle the words that correspond to people: mon village; mon cousin; mon masque; ma mère; mon grand père; mon oncle; mon vélo*");

(c) "vocabulary" category includes questions that ask students, for example, to identify a correct word to complete a sentence (e.g., "*Circle the antonym of the given word: Light—Open; Dark; Sick*");

(d) "grammar" category concerns items around student's knowledge of a sentence structure (e.g., "*Which italicized word(s) represent the nouns of the sentences below … a. [Presumably] the person is sick; b. [The thickness of the] book is 5 cm; c. [The sauce] is less spicy; d. He studied [diligently]*");

(e) "reading ability" category covers items that ask students to read a letter, a sentence, or a paragraph (e.g., "*I want you to read this aloud: Ana is my sister. She plays netball. She is on the school team. Teachers lover her. She is a good player. Was the student able to read the paragraph?*");

(f) "reading comprehension" items typically present students with a paragraph followed by a set of question about the story (e.g., "*Reading Comprehension: Elephant, Frog, Alligator: Who did the alligator see having a bath?*");

(g) "listening comprehension" questions ask students about the story that is read aloud (e.g., "*On Saturday, Lamin and his family stay at home. Mother works in the compound. Father drinks tea with his friend. Binta reads a book. Lamin studies with his friend, Adama. Does Binta play football?*");

(h) "writing" items ask students to write a letter, word, and/or a sentence (e.g., "*Write the letter: T (as pronounced in 'Tomato'), D (as pronounced in 'Donkey'), Dh (as pronounced in 'Adhere')*").

Using this unified framework for content classification, we still find significant variance in content coverage. Figures 2a and 2b show variance in the percent of items for each subdomain across studies for two grade groups—grades 1 through 4 and grades 5 through 8. We find that the range is substantial. Certain test forms consisted of items measuring a single subdomain, e.g., a literacy test with only reading comprehension items or a math test with only algebra questions. Other test forms, on the other hand, had items from each of the subdomains.

In addition to variability in the number of subdomains covered by a single study, the number of items within each of these subdomains varied from relatively equal representation to over-representation by a few subdomains. For example, one study had a relatively equal number of items across the subdomains: 35% number recognition items, 35% subtraction items, and 30% addition items. One of the literacy studies, for example, focused solely on testing grammar knowledge and asked students to fill in missing prepositions (24% of the items) and identify verbs in a sentence (76% of items). In addition to the variation observed in subdomains within grades groups, we noticed that researchers placed a different emphasis on certain types of these subdomains by allocating different numbers of questions. For example, the most common literacy subdomain in grades 1 through 4, word recognition, was measured with just five questions or with 30 questions and more. This pattern holds for each of the subdomains in our sample. That is, what we are measuring effects of programs on, differs substantially from test to test.

3.2.2 Psychometric analysis

We estimated the internal consistency of scores from 211 math and literacy test forms following Cronbach's alpha as defined in equation 1. We restricted this part of the analysis to test forms presented to students at the baseline to avoid learning effects that occur when the same items are included at both baseline and endline. Figure 3 plots the distributions of the reliability estimates across two primary content domains. The alpha coefficient ranges between 0 and 1, with higher values characterizing higher internal consistency. Hence, we would expect a student who is proficient within a content domain to consistently answer questions correctly, whereas a student who has not mastered the materials would be expected to consistently answer questions incorrectly. In our sample, the reliability ranged between .18 and .97, which is consistent with the observation from our systematic review of published studies in the previous section.

We checked reliability estimates reported by standardized tests in the US to put these numbers in perspective. All of these tests report reliabilities around .90 (Reardon and Ho, 2015). For example, PARCC (2017) reports reliability of .91 for literacy and .92 for mathematics, ACT reports reliability of .95 on their literacy test based on a national sample of students who took the ACT in 2018 as part of the state and district testing, or GRE for their 2020 version of the test reports the reliability of .92 for literacy and .95 for mathematics.

We further checked reliability estimates reported for one of the large-scale international assessments, PISA. PISA reports marginal reliabilities, the alternative to Cronbach's alpha within the IRT framework (Adams, 2005), across subdomains due to a complex design of the test forms that rely on more than item response and does not permit computing traditional Cronbach's alphas. These estimates, nevertheless, give us a sense of a range that is considered acceptable within the international community. In PISA 2018, the median values were above 0.80 in all the domains, ranging from .85 to .93 for the computer-based test. While the reliabilities in our sample have a lower degree of internal consistency than those seen in assessments in the US or in leading international assessments, they indicate that the tests typically administered by researchers in development economics are likely to be reliable for population-level inference, which is the goal of most economics research. That is, under classical assumptions, the reliability would increase the variance of test scores and thereby reduce power to detect the average treatment effects, not bias it.

Next, we explored the distributions of the item parameters we estimated following equation 1.2. Figures 4 plots the distribution of the average difficulty parameters per test form, respectively. We find that the distribution of average information is skewed to the right, corroborating the finding that test forms have low reliability. We find that some tests have very low or very high average difficulty parameter estimates. This means that tests are either extremely easy or extremely hard for their targeted population. While focusing on low or high performing students might be helpful for a particular intended purpose (such as, for example, trying to identify students in need of remedial education or, otherwise, gifted program), it does not seem to collect evidence for most students, which is desirable for RCTs with broadly representative populations.

Next, we looked into students' performance on these tests to check for the presence of floor and ceiling effects. We calculated the proportion of students who scored zero on a test (we consider this an extreme indicator of a floor effect) and those who scored 100% by answering all questions correctly (ceiling effect). Figure 5 plots these distributions against each other. On average, 13% of students get zero questions correctly on both literacy and math tests. This contrasts with only 3% of students who managed to answer every question correctly.

We identified several test forms on which nearly 100% of students scored 0. For example, two of such test forms included nine math items in both multiple-choice and constructed-response formats. Figure 6 shows an example of an item from such a test form.

Finally, we explored whether reliability, floor effects, and ceiling effects tend to be higher or lower on average for test forms with a different number of items, a number of subdomains, and administration conditions, controlling for grades. We model them as functions of specific item characteristics following equation 3. Table 2 provides a summary of descriptive statistics of all the variables (panel a) and its estimates on reliability and floor and ceiling effects (panel b), separately for math and literacy and controlling for grades. We see that the relationship between test characteristics, reliability,

and floor/ceiling effects is near zero. There is an interesting positive relationship between the number of content subdomains on literacy tests and reliability ($b = 0.013$, $t(99) = 2.323$, $p < .05$). This relationship is again small but significant, suggesting that more diverse tests have higher internal consistency than homogeneous assessments. Similarly, there is a positive, marginally significant relationship between the presence of orally-administered items on literacy tests and the ceiling effect ($b = 0.021$, $t(99) = 1.821$, $p < 0.1$). That is, tests that have questions administered by proctors are slightly easier, on average, than a written test.

4. Recommendations for practice

Two previous sections highlighted the current state of measurement in developing economics. Here, to illustrate why careful consideration of assessment design is important, we discuss test design and the challenges we faced in several previous studies that we were involved in. The reason for focusing only on our own previous studies is simply that this is work where we are best placed to comment on both the objectives of test design and the constraints that we faced in practice, which inform the eventual design used. We then present a general set of recommendations for researchers on test design and field procedures in a variety of contexts.

4.1 Case study: Muralidharan, Singh and Ganimian (2019)

The Mindspark study provides a good example of a typical impact evaluation in developing countries (Muralidharan *et al.*, 2019). For this study we recruited 619 students from middle schools in Delhi who were randomly assigned to either receive a supplementary computer-aided instruction program or to be in the control group. The primary outcome for the study was student achievement in mathematics and Hindi, which we measured using independent assessments both at a baseline and an endline.

Even this simple set-up, however, raises numerous challenges. The most significant of these related to the difficulty level of the test. Put bluntly, should our assessments target the difficulty level that (a) students are supposed to learn in a particular grade, (b) the achievement level they are actually at or (c) the level at which the intervention aims to have an effect? In many LMIC populations, the median student is substantially behind grade level achievement, the range of achievement in a single grade can be large, and the interventions may only target students at particular level of achievement. The distinction between these three possible levels can be particularly stark at the middle school level. Designing a test that is informative at all three levels is often difficult, especially in settings of large heterogeneity, and difficult to resolve on conceptual and statistical grounds. The Mindspark intervention focused on delivering personalized instructions for students at their level of proficiency—however, our sample spanned multiple grades, and therefore our tests had to also be informative across the full range of achievement distribution.

This focus on a broad range of skills was distinct from the local curriculum, both instruction and assessments in school, which are the signals most students and parents receive about academic achievement. While the academic literature and the policy discussions around school reforms privilege measures of student competence, such as the ability to read with comprehension and complete basic arithmetic computation, the tests that schooling systems themselves administer frequently privilege rote memorization and the recall of facts (Burdett, 2017). These school-based tests also often have high stakes: many education systems mandate grade repetition if students fail to attain a passing grade in end-of-year exams, and the high-stakes exams that students take at particular education milestones are used for selective admissions and also as signals to certify the skills of students. Given that economists' interests in education is often motivated by the effects of education on long-term life outcomes, it is not obvious which of these we should privilege in the measurement of student achievement. Accordingly, we also collect student scores from school tests. These choices are consequential: On the school exams, it appears that only the top third of students benefit from the intervention, but this is only because the intervention personalized instruction to students' achievement levels: the remaining students did make substantial improvements, but this does not show up on school exams since these improvements are on skills substantially below the curriculum-mandated levels.

In addition to the large variation in difficulty and content coverage, we wanted to bring our baseline and endline assessments on the same scale: so, at the time of designing the endline assessments, we kept a substantial share of common items from the baseline, retaining especially those items which were highly discriminatory at all levels of proficiency at the baseline. As a result, students in the sample, mostly in Grades 6-9, had grade-specific booklets with a substantial degree of common items across booklets. We were able to achieve both of our goals with Item Response Theory (IRT) approach, which underpins all large-scale international assessments, as well as tests in the US (Jacob and Rothstein, 2016). This approach is not yet common in the development economics literature.

In this study, we relied on the 3PL IRT model described above (see equation 2) to score tests that is most commonly used for multiple-choice questions. We landed with the test items in the assessments that spanned a very wide range of difficulty. For example, in mathematics, at the lowest end of ability, items merely required a knowledge of counting, while at the high end, students were asked to interpret complex data chart based on publicly released items from international large-scale assessments, such as PISA and TIMSS. This was crucial for recovering smooth distributions of student achievement without ceiling or floor effects in the test, and for us to capture treatment effects across the ability distribution and for making the baseline tests informative and predictive, which, in a small trial, helped substantially with statistical power.

The importance of a wide range of items, both in ability and content-coverage, also helped with substantial supplementary analyses beyond the main treatment effects on aggregate scores: thus, we could provide effects not just on the linked IRT scores but also on (a) specific subject-specific competences, (b) on test items that were or were not in EI-designed assessments and (c) on test items which were testing skills at or below grade-level. All these features together, which were only made

possible by ex-ante attention to test design, helped characterize treatment effects in much more detail than in many settings where the investments in terms of fieldwork costs and data collection are quite similar.

Since the comparability is a big issue in the current practice, as we illustrated in Sections 2 and 3, we concerned ourselves with ways we can build comparability in our assessments. One option, of course, is to administer exactly the same test. A single test comprised of exactly the same items is, however, unlikely to be feasible or appropriate as a general solution. For instance, a student in Grade 4 and another in Grade 8 may need to be tested on different competences, even to judge effect sizes of the same underlying intervention (say, a scholarship of equivalent monetary value) in the same setting. A more satisfactory (although model-based and analytically complex) alternative is to use IRT, as we described above.

A final consideration, which we alluded to before and is often ignored, relates to the fact that student ability has no natural metric. Test scores are not interval scaled and only express such ability on an ordinal scale: as such, any rank-preserving transformation could be equally valid as a test score. This ordinality of aggregate test scores renders many analyses of test scores, and interpretations of resulting estimates, suspect especially concerning changes in achievement and inter-group differences in achievement. That said, this issue is not specific to developing countries and affects nearly all quantitative research using test scores across disciplines.

4.2 Recommendations

Our aim in this paper is not just to highlight considerations around test design that we think are important for impact evaluations in developing countries, but also to suggest some guidance for practice. Thus, in this section, we focus on recommendations for researchers at all stages from test design to eventual analysis.

Our recommendations for item selection and test design have been interspersed already in the previous sections. Tests should be designed to measure a broad range of proficiency, there should be a clear mapping of content at the item level to subdomains, and ideally items should come from multiple sources and allow for linking comparably with other samples and over time. One further note of caution in this regard relates to making sure that the test lends itself easily to generating a smoothly distributed aggregate score: such a summary metric at the subject level is what most economics-focused evaluations need. While this is the case for most assessments that researchers might design and use, it is notably not the case for assessments such as the Early Grade Reading Assessment and the short ASER tests.

Our central recommendation for test design though is, wherever possible, to pilot the assessments in the context that they will be administered and to do this with a larger item bank than is intended for final use. Specifically, even at the scale of a few classrooms, pilot test data allows researchers

to ensure that the test distribution does not suffer from floor or ceiling effects, that selected items have some variation across individuals, that items are positively correlated with each other and that the test is internally coherent with sufficiently high reliability, which can be easily checked with Cronbach's alpha. If the design features any psychometric linking, whether across test forms or over time, it is well worth doing the linking with the pilot data itself to make sure that linked items do not suffer from Differential Item Functioning and are well-distributed across the range of proficiency distribution. In rare circumstances when we have needed to roll out assessments without piloting (typically due to very short windows in which to design and field a baseline), we have inevitably regretted the inability to refine our assessments: even across different Indian states, which share a common curriculum framework, heterogeneity across samples frequently means that tests that have high discrimination in, say, Andhra Pradesh are much less informative in other settings.

As we have documented throughout this paper, these details of testing matter. Thus, it is important that researchers tell us, in their papers, about the assessments they used—what was tested, how it was designed and administered, how the tests were scored and scaled, and what the psychometric properties of the aggregate test scores are. Without these details, assessing the validity of researchers' interpretations of obtained effect sizes is hard. In most cases it is very informative to have details of testing, and the distribution of aggregate test scores at least in a data appendix. If possible, researchers should also archive item-level data and not just the aggregate test scores (as done most often, including by us, when uploading data to journal websites).

These recommendations also extend to analysis, although that is not the primary focus of this article. For better or worse, disciplinary norms all focus on standardized effect sizes on test scores and so, presenting these is important (and even more informative if the aggregate test distributions are shown). But, as has been pointed out in this paper and elsewhere, this is rarely enough and standard deviations not comparable across samples. Thus, we recommend not just presenting the effect on aggregate test scores but also on specific competences and subdomains. When expressed as the effect of the treatment on the probability of successfully answering these specific types of questions, it provides an external benchmark that is much more easily understood and explained to external audiences. It may also be worthwhile to express treatment effects in relation to other meaningful magnitudes in the context, such as the SES gap or the magnitude of learning under business-as-usual in that setting. Note, though, that any such expressions only serve to improve exposition and benchmarking within contexts, not the (lack of) comparability: clearly, just as a standard deviation can vary across contexts, so can learning over time in the absence of interventions (see e.g, Singh, 2020).

5. Conclusions

This paper has surveyed current practice in the economics of education in developing countries and discussed ways to make the underlying metrics on which this literature is based sound and more comparable. Our discussion targets individual researchers who design their own studies and

assessments, possibly in populations where validated assessments are unavailable, and who are principally interested in the estimation of treatment effects (although many of the recommendations would also hold true for doing descriptive research in these settings).

The key finding for this work relates to its limitations. Our systematic review showed the lack of documentation of the properties of learning outcomes. Even when attempting to evaluate tests post hoc, we found that only 9.5% of studies archived item-level replication data files. Second, content analysis of question wordings from item-level data we accumulated revealed substantial variation in content coverage of the skills tested, even when students of similar grades are being tested in similar subjects. For instance, the IQR for the number of distinct subdomains on a single math test varied from 5 to 13. Finally, when exploring psychometric properties of the item-level data, we found (a) reliability to vary substantially across studies (.18 to .97 range); (b) test characteristics, such as the number of items and subdomain on a test form, had near-zero relationship with reliability and the presence of floor and ceiling effects across math and literacy; and (c) tests comprised of multiple-choice items showed 21% probability of getting the correct answer by chance.

Again, the list of studies included in the systematic review was not an exhaustive enumeration of all papers that used test scores in development economics. We believe that this list of recent high-quality studies is adequate to characterize current practices in the economic literature. We focused primarily on studies that used experimental designs with independent data collection. That excluded, for instance, work relying on panel data methods with administrative data or most work relying on large nationally representative repeated cross-sectional surveys such as the ASER datasets in India. We made this choice for two reasons. First, such studies comprise a substantial portion of the academic literature focused on improving student achievement and are, therefore, allow us to outline the current state of practice, the critical goal of this exercise. Second, most researchers using data from public use datasets, whether from administrative sources or large-scale assessments, such as PISA or TIMSS or civil society-led assessments in South Asia or East Africa, have no opportunity to influence the design or administration of these tests.

Stand-alone randomized controlled trials (RCTs), with the potential for tying measurement closely to intervention design and the samples in which the intervention is being administered (which may themselves be unrepresentative, see Allcott, 2016), have the most to gain from improving test designs, and our findings hint that way. Moving forward, one of our goals should be to achieve comparability in interpreting effect sizes from different studies. To reach it, we may want to engage in establishing a common formal standard for reporting and analysis of measures of learning outcomes, given the degree of control researchers have on study design and the extent to which such standards already dictate substantial parts of analysis and research practice for RCTs (such as AEA trial registry and pre-analysis plans).

References

American Educational Research Association *et al.* (2014). *Standards for educational and psychological testing*. Amer Educational Research Assn.

— *et al.* (2018). *Standards for educational and psychological testing*. American Educational Research Association.

Barrera-Osorio, F. and Filmer, D. (2016). Incentivizing schooling for learning: Evidence on the impact of alternative targeting approaches. *Journal of Human Resources*, **51** (2), 461–499.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, **38** (4), 295–317.

Burdett, N. (2017). Review of high stakes examination instruments in primary and secondary school in developing countries. *Research on Improving Systems of Education*, **17** (018), 18–21.

Connolly, P., Keenan, C. and Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, **60** (3), 276–291.

Churchill Jr, G. A. and Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, **21**(4), 360–375.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16** (3), 297–334.

Filmer, D. and Schady, N. (2014). The medium-term effects of scholarships in a low-income country. *Journal of Human Resources*, **49** (3), 663–694.

Glewwe, P. and Muralidharan, K. (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the Economics of Education*, vol. 5, Elsevier, pp. 653–743.

Grosh, M. and Glewwe, P. (2000). *Designing household survey questionnaires for developing countries*. Washington, DC: World Bank.

Jacob, B. and Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives*, **30** (3), 85–108.

Jaju, A. and Crask, M. R. (1999). The perfect design: optimization between reliability, validity, redundancy in scale items and response rates. In *American Marketing Association. Conference Proceedings* (Vol. 10, p. 127). American Marketing Association.

Koch, A., Nafziger, J. and Nielsen, H. S. (2015). Behavioral economics of education. *Journal of Economic Behavior & Organization*, **115**, 3–17.

Kolen, M. J. and Brennan, R. L. (2014). Nonequivalent groups: Linear methods. In *Test equating, scaling, and linking*, Springer, pp. 103–142.

Kremer, M., Brannen, C. and Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, **340** (6130), 297–300.

Loyalka, P., Popova, A., Li, G. and Shi, Z. (2019). Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics*, **11** (3), 128–54.

Muralidharan, K. (2017). Field experiments in education in developing countries. In *Handbook of economic field experiments*, vol. 2, Elsevier, pp. 323–385.

—, Singh, A. and Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in india. *American Economic Review*, **109** (4), 1426–60.

Reardon, S. F. and Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, **40** (2), 158–189.

Sandefur, J. (2018). Internationally comparable mathematics scores for fourteen African countries. *Economics of Education Review*, **62**, 267–286.

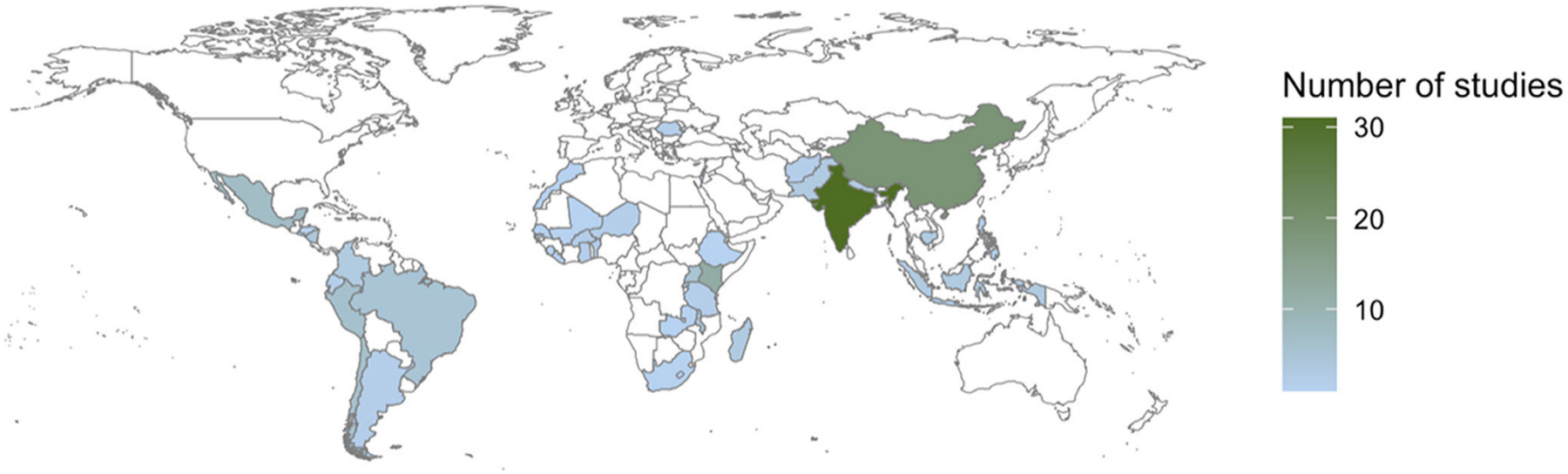
Singh, A. (2020). Learning more with every year: School year productivity and international learning divergence. *Journal of the European Economic Association*, **18** (4), 1770–1813.

van der Linden, W. J. and Hambleton, R. K. (Eds.). (2013). Handbook of modern item response theory. Springer Science & Business Media.

World Bank, W. (2017). *World development report 2018: Learning to realize education's promise*. The World Bank.

Figures and tables

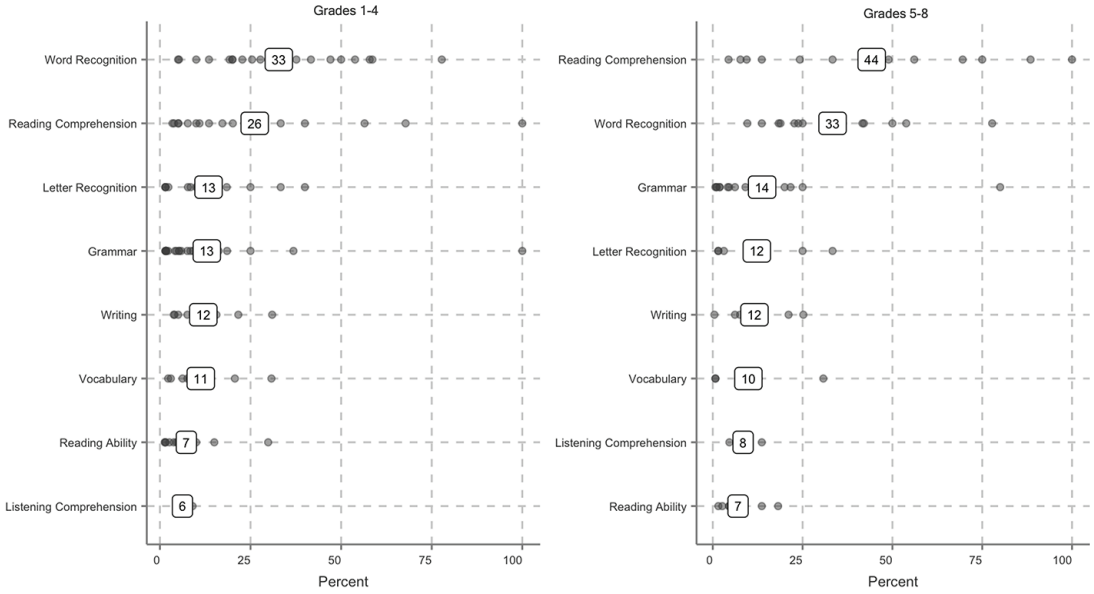
FIGURE 1. Country representation in systematic review



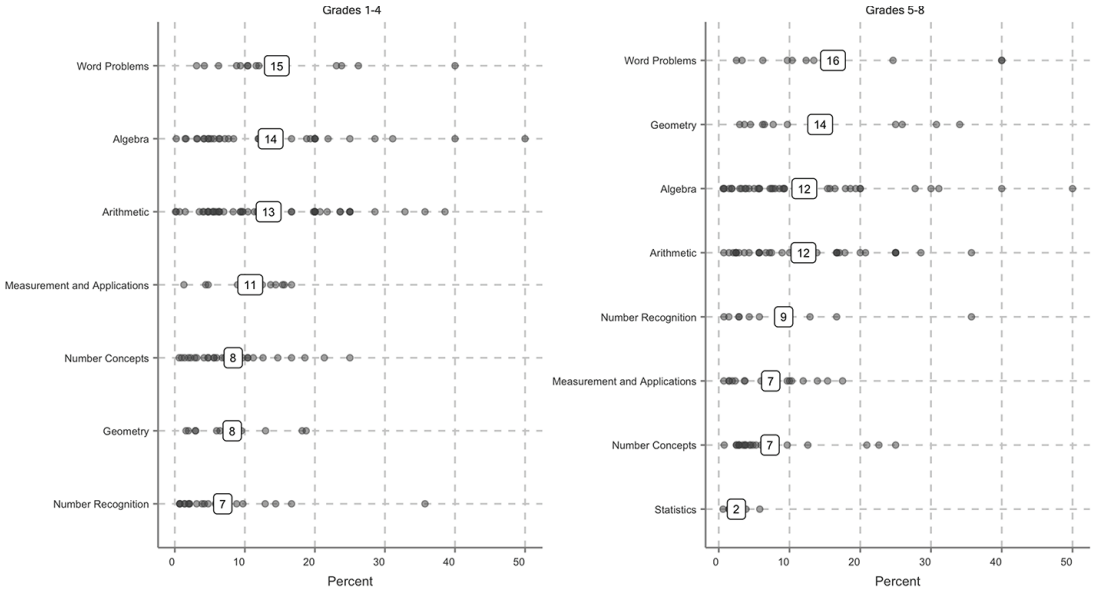
Note: The review covered 158 studies. Darker color indicates a larger number of studies conducted in a corresponding country (e.g., 32 studies were conducted in India vs. one study in South Africa).

FIGURE 2. Subdomains covered in test forms in literacy and math

(a) Literacy subdomains

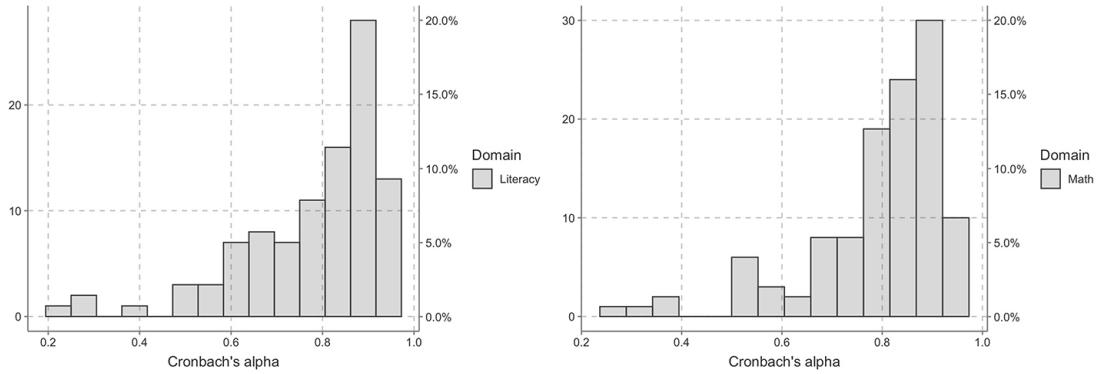


(b) Math subdomains



Note: This figure presents the distribution of competences in the Item Bank collated from various studies, as described in Section 3.

FIGURE 3. Distribution of reliability estimates in math and literacy



Note: This figure presents Cronbach's alpha in math (left panel, N = 102 test forms) and literacy (right panel, N = 109 forms) in the review.

FIGURE 4. Item difficulty and test information across subdomains

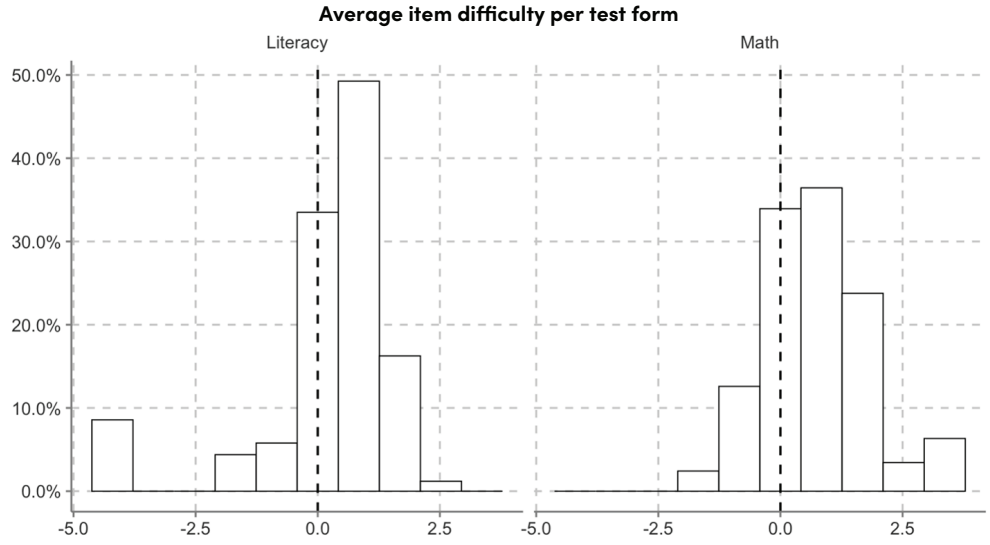
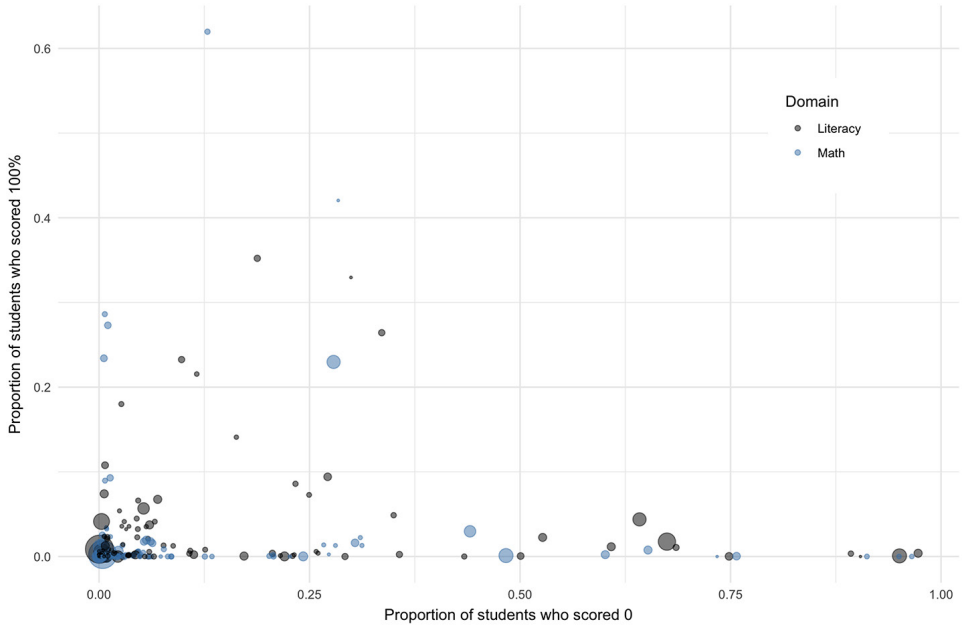


FIGURE 5. Presence of floor and ceiling effects



Note: Proportion of students scoring a zero vs proportion of students scoring 100% weighted by the sample size and split by domains.

FIGURE 6. Sample math item from a test form with a 96.5% of the floor effect

Find the value of exterior angle X in the following figure.

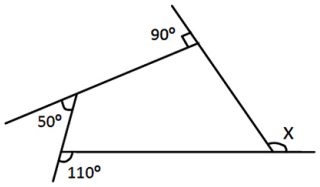


TABLE 1. The systematic review sample (top panel) and researcher-designed assessments' characteristics (bottom panel)

				Assessment Characteristics		
		Count	Percent		Count	Percent
Full sample (N = 158)	**Region**			**Test(s)**		
	African	46	29.1%	achievement test only	56	35.4%
	Eastern Mediterranean	3	1.9%	achievement test and behavioral outcome	66	41.8%
	European	2	1.3%	achievement test and psychological outcome	7	4.4%
	Middle Eastern	1	0.6%	other	29	18.4%
	Americas	41	25.9%	**Test origin**		
	South-East Asia	65	41.1%	government, state, or district exam	51	32.3%
	Urbanicity			vendor-designed	22	13.9%
	Urban	25	16.9%	teacher-designed	3	1.9%
	Rural	51	34.5%	researcher-designed	64	40.5%
	Rural and urban	59	39.9%	mix: researcher-designed and else	9	5.7%
	Nationally representative sample	20	13.5%	not reported	9	5.7%
	Not reported	3	2.0%	**Domain**		
	School type			mathematics (only)	18	11.4%
	Public (only)	71	48.0%	literacy (only)	8	5.1%
	Private (only)	12	8.1%	science (only)	0	0.0%
	Public and private	40	27.0%	cognition (only)	4	2.5%
	Other	9	6.1%	mathematics and literacy	86	54.4%
	Not reported	26	17.6%	mix: two or more of the above	37	23.4%
	Grades			not reported	5	3.2%
	Elementary	78	52.7%			
	Middle school	14	9.5%			
	Elementary and middle	17	11.5%			
	High school	26	17.6%			
	Mix of all	21	14.2%			
	Not reported	2	1.4%			

TABLE 1. (Continued)

		Assessment Characteristics				
		Count	Percent		Count	Percent
Researcher-designed tests (N = 73)	**Administration condition**			**Domain**		
	at school	36	48.6%	mathematics (only)	12	16.2%
	at home	17	23.0%	literacy (only)	4	5.4%
	both	3	4.1%	science (only)	0	0.0%
	not reported	17	23.0%	cognition (only)	0	0.0%
	Mode of administration			mathematics and literacy	38	51.4%
	written	16	21.6%	mix: two or more of the above	17	23.0%
	oral	2	2.7%	not reported	2	2.7%
	both	13	17.6%	**Math subdomain**		
	not reported	42	56.8%	number properties	10	13.5%
	Alignment to curriculum			algebra	1	1.4%
	to district or country	45	60.8%	multiple subdomains	15	20.3%
	not aligned	9	12.2%	not reported	47	63.5%
	not reported	19	25.7%	**Literacy subdomain**		
	Borrowed items			early literacy	4	5.4%
	yes	32	43.2%	vocabulary/word analysis	4	5.4%
	no	36	48.6%	reading fluency	1	1.4%
	not reported	5	6.8%	reading comrehension	1	1.4%
	Reported number of items	21	28.8%	multiple subdomains	16	21.6%
	Reported duration of tests	15	20.6%	not reported	47	63.5%
				Reported reliability	3	4.1%
				Range: 0.65—0.90		

TABLE 2. Quality of a test form and various item characteristics: descriptive statistics of the variables

	Literacy (N = 109)	Math (N = 102)	Overall (N = 211)
Reliability			
Mean (SD)	0.822 (0.151)	0.822 (0.123)	0.822 (0.138)
Median [Min, Max]	0.875 [0.179, 0.963]	0.856 [0.287, 0.971]	0.858 [0.179, 0.971]
Floor effect			
Mean (SD)	0.142 (0.228)	0.121 (0.215)	0.132 (0.221)
Median [Min, Max]	0.0432 [0, 0.973]	0.0240 [0, 0.965]	0.0351 [0, 0.973]
Ceiling effect			
Mean (SD)	0.0297 (0.0636)	0.0268 (0.0878)	0.0283 (0.0761)
Median [Min, Max]	0.00384 [0, 0.352]	0.000511 [0, 0.620]	0.00224 [0, 0.620]
Number of items			
Mean (SD)	31.9 (25.1)	32.6 (21.8)	32.2 (23.5)
Median [Min, Max]	27.0 [4.00, 125]	30.0 [4.00, 137]	28.0 [4.00, 137]
Number of subdomains			
Mean (SD)	4.75 (2.56)	8.39 (3.48)	6.47 (3.53)
Median [Min, Max]	5.00 [1.00, 11.0]	8.00 [3.00, 24.0]	6.00 [1.00, 24.0]
Orally-administered items			
Mean (SD)	0.495 (0.502)	0.179 (0.385)	0.347 (0.477)
Median [Min, Max]	0 [0, 1.00]	0 [0, 1.00]	0 [0, 1.00]
Grades			
Kindergarten or lower	8 (7.3%)	5 (4.9%)	13 (6.2%)
Grades 1–4	68 (62.4%)	56 (54.9%)	124 (58.8%)
Grades 5–8	30 (27.5%)	33 (32.4%)	63 (29.9%)
Grades 9+	3 (2.8%)	8 (7.8%)	11 (5.2%)

TABLE 3. Estimated effects of item characteristics on test quality, controlling for grades

	Dependent Variable:					
	Reliability	Floor Effect *Math*	Ceiling Effect	Reliability	Floor Effect *Literacy*	Ceiling Effect
	(1)	(2)	(3)	(4)	(5)	(6)
Number of items	0.004***	−0.006***	−0.001**	0.001	−0.003***	−0.0004**
	(0.001)	(0.002)	(0.0004)	(0.001)	(0.001)	(0.0002)
Number of subdomains	0.001	0.001	−0.002	0.013**	0.001	0.0005
	(0.004)	(0.007)	(0.002)	(0.006)	(0.008)	(0.002)
Orally-administered items	0.010	−0.002	0.011	0.025	0.0002	0.021*
	(0.021)	(0.040)	(0.018)	(0.023)	(0.046)	(0.012)
Grades 5–8	−0.078***	0.095*	0.023	−0.030	−0.047	0.016
	(0.028)	(0.054)	(0.021)	(0.031)	(0.045)	(0.016)
Grades 9+	−0.137**	−0.104**	−0.016*	−0.368**	−0.147***	−0.015**
	(0.053)	(0.040)	(0.009)	(0.152)	(0.047)	(0.007)
Kindergarten or lower	−0.038	0.362***	−0.009	0.040	0.278**	−0.029***
	(0.045)	(0.134)	(0.009)	(0.029)	(0.137)	(0.009)
Constant	0.734***	0.253***	0.055***	0.735***	0.245***	0.027**
	(0.047)	(0.057)	(0.019)	(0.044)	(0.061)	(0.010)

Note: Robust standard errors in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01. The tables show the relationship between test quality and certain items and student characteristics across math (columns 1 through 3) and literacy (columns 4 through 6) domains and its descriptive statistics (top panel). The dependent variables are: reliability is a measure of internal consistency of a test measured with Cronbacha's alpha (see equation 1); floor effect is the proportion of students who answered zero questions correctly on a test; and ceiling effect, in turn, is the proportion of students who answered all questions correctly on a test. Number of items is the number of items on a test form. Number of subdomains is the number of unique subdomains questions on a test form have been reflective of. Similarly, the orally-administered items variable indicates whether the test had questions that have been orally-administered by proctors rather than have been presented to students in a written form.

Appendix

Codebook

- Report and Setting

1. Study—ID
2. Reference
3. Author (break down the APA citation and include only authors (e.g., Behrman, J.R., Parker, S.W., Todd, P.E., 2009. Medium-term impacts of the Oportunidades conditional cash transfer program on rural youth in Mexico. In: Klasen, S., Nowak Lehmann, F. (Eds.), Poverty, Inequality and Policy in Latin America. MIT Press, Cambridge, MA. ➔ Behrman, Parker, Todd)
4. Pub—year: Publication year
5. Rep—type: Report type
   - 1 = Journal article
   - 2 = Book or book chapter
   - 3 = Dissertation
   - 4 = MA thesis
   - 5 = Private report
   - 6 = Govt report (state, federal, or district)
   - 7 = Conference paper
   - 8 = Other
6. Peer—rev: Was the report peer-reviewed?
   - 1 = Yes
   - 0 = No
7. Org: What type of organization produced this report?
   - 1 = University
   - 2 = Govt entity (specify)
   - 3 = Contract research firm (specify)
   - 4 = Other
   - 5 = joint project
8. Funder
   - 1 = Gov entity
   - 2 = Private foundation
   - 3 = Other
   - 4 = Mix
9. Funder = J-PAL
   - 1 = Yes
   - 0 = No

10. Are data publicly available
    - 1 = Yes
    - 0 = No
11. Country
12. Region (specify regions, if applicable)
13. Urban: Geographic region/Urbanicity
    - 1 = urban
    - 2 = suburban
    - 3 = rural
    - 4 = mix
    - 5 = nationally representative sample
14. Rep—sample: Was the sample representative (get directly from previous code; set filter if "prev column" = 5, 1, 0)
    - 1 = yes
    - 0 = no
15. FirstYear: First year of data collection
16. LastYear: Last year of data collection
17. Dur: Duration of the study in months

- Participants

18. N: Number of participants
19. PercFem: Percent female
20. Attrition: Is there evidence that the number of students pre-tested is higher than the number of students post-tested (sample attrition)?
    a) 1 = Yes b) 0 = No c) NA
21. PercAttr: If "attrition" = 1, what percentage (if specified)?
22. Age
23. Grades
    - 1 = Early elementary (grades K-2)
    - 2 = Upper elementary (grades 3–5)
    - 3 = mix of elementary grades
    - 4 = Middle school
    - 5 = mix of elementary and middle school
    - 6 = High school
    - 7 = mix of elementary and high
    - 8 = mix of all
24. SchType: Type of school students attended:
    - 1 = public
    - 2 = private
    - 3 = other; specify

25. SESReport: Was there any information on sample SES, if yes, specify (e.g., using some SES scale such as SLI or anything else, e.g., annual income), if not = 0
26. SESstatus: Sample was predominately (50% or greater)
    - 1 = low income
    - 2 = middle income
    - 3 = high income
    - 4 = mix
    - NA
27. LowSES: Low SES (set as "if, then" function from previous question)
    - 1 = Yes
    - 0 = No
28. Disability: Were students with disability part of the study
    - 1 = Yes
    - 0 = No
    - 2 = NA

- Intervention design Effect sizes

29. Experiment: Was this an experimental study, i.e., RCT? a) 1 = Yes b) 0 = No
30. Design-type: Which design type was implemented (e.g., block design, SMART)? a) 1 = RCT with pretest data b) 2 = RCT without pretest data c) 3 = non-equivalent control (NEC), a priori matching on achievement and demographic variables d) 4 = RDD e) 5 = DID f) 6 = add additional if come across
31. ITT (were intent-to-treat analysis conducted) a) 1 = Yes b) 0 = No
32. TOT (were treatment on the treated analysis conducted) a) 1 = Yes b) 0 = No
33. Tx-mean: treatment group posttest mean
34. C-mean
35. Tx-sd: treatment group standard deviation
36. C-sd
37. N-tx (number of participants in treatment group)
38. N-c (number of participants in control group)
39. Pooled-sd
40. D: Study effect size (in sd units)
41. D-var: effect size variance
42. St-s.e.: Study standard error
43. St-e.s.: Study effect size for male (if applicable)
44. s.e. for males
45. Study effect size for female (if applicable)
46. s.e. for females
47. Study effect for other subgroups (add extra columns as needed)
48. s.e. for other subgroups

- Outcome

1. Baseline-true: Does study have a baseline test?
   - 1 = Yes
   - 0 = No
2. Endline-true: Does study have endline test?
   - 1 = Yes
   - 0 = No
3. Pilot-true: Was assessment piloted/pretested with the similar group of students?
   - 1 = Yes
   - 0 = No
4. Outcome
   - 1 = student achievement test only
   - 2 = student ach test behavioral measure
   - 3 = stud ach test psych outcome (e.g., cog ability)
   - 4 = other (specify)
5. Outcome-describe: Provide information about the assessment (0 = if only say something like "math was the outcome of interest")
6. Standardized-test
   - 1 = standardized: governmental, state, or district
   - 2 = standardized: vendor
   - 3 = teacher-designed
   - 4 = researcher-designed
   - 5 = other, specify
7. Common-items: Were some of the questions borrowed from existing assessments?
   - 1 = Yes
   - 0 = No
8. If "common-items" = 1, specify
9. Base-equiv: Were groups' pretest academic scores statistically equivalent
   - 0 = No
   - 1 = Yes
   - 2 = scores were matched
   - 3 = didn't give pretest
   - 4 = not reported
10. Tx-higher: If the pretest scores for tx and control groups were not statistically equivalent, did the tx group have a higher mean score?
    - 1 = Yes
    - 0 = No

11. Base-iden: If a base/pretest was given, was it identical to the post-test?
    - 0 = neither eqiv nor identical
    - 1 = identical
    - 2 = equivalent
    - 3 = pretest was prev year's state test, unsure if vertically linked
12. Posttest-date: How long after the end of the program were outcomes for tx and c groups measured?
    - 1 = 0–1 months
    - 2 = 1–2 months
    - …
    - 13 = 12–13 months
    - 14 = 14 or more months
13. Aligned-goals: Was the outcome measure aligned to the goals of the program?
    - 1 = Yes
    - 0 = No
14. Aligned-curric: Was the outcome measure aligned to the curriculum?
    - 0 = No
    - 1 = Yes, to local curriculum
    - 2 = Yes, to country
    - NA
15. Norm-ref: Was the assessment norm-referenced (reported in grade equiv score)?
16. Scoring: [Need to figure out what's the best way to reflect the practices; ignore for now]
17. Subjective: Does the assessment involve subjective decision-making (e.g., scoring on a rubric)?
    - 1 = Yes
    - 0 = No
18. Raw-score: If tests were given, how were scores reported?
    - 0 = not reported
    - 1 = raw scores
    - 2 = scaled measurement (scaled IRT scores, Bayes Nets)
    - 3 = standardized
    - 4 = normalized
    - 5 = percent correct
    - 6 = neither (coefficient, etc.)
    - 7= other, specify
19. If normalized/standardized score was used, describe how it was normalized (put NA if authors do not report)

20. Reliability (if give range, use lowest value)
    - 0 = not reported
    - 1 = reported less than .7
    - 2 = reported, .71–.8
    - 3 = reported, .81–.89
    - 4 = reported, equal or ¿ .9
21. Domain
    - 1 = mathematics
    - 2 = literacy
    - 3 = science
    - 4 = cognitive ability
    - 5 = geography/history/civics/religion (GHCRE)
    - 6 = home science/business education (HS-BE)
    - 7 = mix: math lit
    - 8 = mix: math cog
    - 9 = mix: other
    - 10 = other
22. If "domain tested" = 2, what form of literacy has been tested
    - 1 = unspecified
    - 2 = English
    - 3 = Native language
    - 4 = Eng Native
23. If "form of literacy" = 3 and 4, specify the language(s)
24. Sub-dom-math: Subdomain math
    - 1 = Number properties (i.e., arithmetic operations, including addition, subtraction, multiplication, division)
    - 2 = Measurement (attributes such as length, perimeter, distance, height, weight/mass, time, temperature; units, such as inch, pound, hour, cm, liter, gram)
    - 3 = Geometry (calculating lengths, areas, volumes, common shapes; familiarity with plane (lines, circles, triangles, squares) and space (cubes, spheres, cylinders))
    - 4 = Data analysis, statistics, and probability (collecting, organizing, summarizing, interpreting data, describing center, spread, shape. E.g., knowing some numbers being able to pose a question that can be answered with data)
    - 5 = Algebra (functions, equations, tables, graphs, proportionality and rate)
    - 6 = Mix
    - 7 = All
    - 8 = other

25. Sub-dom-lit: Subdomain literacy
    - 1 = early literacy (phonemic awareness, phonics, sight words)
    - 2 = vocabulary/word analysis
    - 3 = reading fluency
    - 4 = reading comprehension
    - 5 = writing
    - 6 = language
    - 7 = spelling
    - 8 = mix
    - 9 = other, specify
26. Math-real: Were math questions based on real-world examples? For example, joining two collections or laying two lengths end to end can be described by addition, whereas the concept of rate depends on division.
    - 0 = No
    - 1 = Yes
    - 2 = Mix
    - 3 = Unknown
27. Lit-real: Were literacy questions based on real-world examples?
    - 0 = No
    - 1 = Yes
    - 2 = Mix
    - 3 = Unknown
28. Items-avail: Do researchers present items either in append or elsewhere?
    - 1 = Yes
    - 0 = No
29. Tech-report: Does assessment have a tech report?
    - 1 = Yes
    - 0 = No
30. Admin-cond: Where was assessment administered?
    - 1 = at school
    - 2 = at home
31. Mode: mode of administration
    - 1 = written, paper and pencil
    - 2 = written, DBA (digitally-based assessment)
    - 3 = oral
    - 4 = other, explain
32. Length: Length of the assessment in minutes

33. Diff-base: Difficulty of the assessment at baseline (describe if authors provide any information or if they report score distributions)
34. Diff-end: Difficulty of the assessment at endline (describe if authors provide any information or if they report score distributions)
35. Floor-ceiling: Was floor or ceiling effect present based on the description and score distributions?
    - 0 = No
    - 1 = Yes, floor at baseline
    - 2 = Yes, floor at endline
    - 3 = Yes, ceiling at baseline
    - 4 = Yes, ceiling at endline
    - 5 = mix of "yes"
    - 6 = unknown
36. N-items-base: Number of items at baseline
37. N-math-base: Number of items at math baseline
38. N-lit-base: Number of items at literacy baseline
39. N-sci-base: Number of items at science baseline
40. N-items-end: Number of items at endline
41. N-math-end: Number of items at math endline
42. N-lit-end: Number of items at literacy endline
43. N-sci-end: Number of items at science endline
44. Item-type: Types of items
    - 1 = multiple-choice
    - 2 = open/constructed responses
    - 3 = composition/essay
    - 4 = 1 2
    - 5 = 1 3
    - 6 = 2 3
    - 7 = all
    - 8 = other, specify
45. Weights: Were weights assigned when building a composite score (e.g., a composite of reading and math tests)?
    - 1 = Yes
    - 0 = No
46. If "weights" = 1, specify
47. Stakes: Assessment stakes for student
    - 1 = low-stakes (no personal consequences for students)
    - 2 = high-stakes (decisions will be made based on the performance, such as grade retention, remedial education)