Aligning Test Scoring Procedures with Test Uses of the Early Grade Mathematics Assessment: A Balancing Act

Leanne R. Ketterlin-Geller
Southern Methodist University

Lindsey Perry
Southern Methodist University

Linda M. Platas
San Francisco State University

Yasmin Sitbakhan
RTI International

Abstract

Test scoring procedures should align with the intended uses and interpretations of test results. In this paper, we examine three test scoring procedures for an operational assessment of early numeracy, the Early Grade Mathematics Assessment (EGMA). The EGMA is an assessment that tests young children's foundational mathematics knowledge and has been administered in more than 25 countries. Current test specifications call for subscores to be reported for each of the eight subtests on the EGMA; however, in practice, composite scores have also been reported. To provide users with empirically-based guidance on the appropriateness and usefulness of different test scoring procedures, we examine the psychometric properties – including the reliability and distinctiveness of the results – and usefulness of reporting test scores as (1) total scores, (2) subscores, and (3) composite scores. These test scoring procedures are compared using data from an actual administration of the EGMA. Conclusions and recommendations for test scoring procedures are made. Generalizations to other testing programs are proposed.

Keywords

Early Grade Mathematics Assessment, EGMA, test scoring procedures, testing programs

Introduction

The purpose of this paper is to examine test scoring procedures for the Early Grade Mathematics Assessment (EGMA) operational testing program and determine the approach that is psychometrically appropriate and useful. The EGMA tests young children's foundational mathematics knowledge in a series of eight subtests. It is typically administered to students in Grades 1-3 to determine their basic number concepts and facility with operations and applied arithmetic.

EGMA results are primarily used by researchers and policy makers as the dependent measure for program evaluation purposes.

Corresponding Author:
Leanne R. Ketterlin-Geller, Simmons School of Education and Human Development, Southern Methodist University, PO Box 750114, Dallas, TX 75275-0114
Email: lkgeller@smu.edu

 Citation: Ketterlin-Geller, Leanne R., Perry, Lindsey, Platas, Linda M. & Sitbakhan, Yasmin (2018). Aligning test scoring procedures with test uses of the early grade mathematics assessment: A balancing act. *Global Education Review, 5* (3), 143-164.

The results from the EGMA provide stakeholders with data that can guide reforms in policies and practices, and inform intervention design and evaluation (Platas, Ketterlin-Geller, & Sitabkhan, 2016). Baseline measurement of children's skills on the EGMA informs prospective reforms in content standards, benchmarking, and teacher education programs. Interventions with pre- and post-measurements can include curricula, classroom practices and materials, teacher education and training, coaching models, textbooks, and combinations of these elements. To facilitate these decisions, the developers of the EGMA recommend that results from each subtest be reported individually as subscores (RTI International, 2014), as opposed to aggregating scores from multiple subtests to form a composite or total score. This is the most common practice for reporting EGMA results (c.f., Brombacher et al., 2015; Piper & Mugenda, 2014; Torrente et al., 2011).

While useful in many ways, subscore reporting has some limitations and has generated controversy in the measurement field (Sinharay, Haberman, & Puhan, 2007). Subscores may not support all of the users' desired decisions, leads to lengthy reports and presentations of results, and may be difficult to interpret for individuals who are not experts in early grade mathematics. For example, if policy makers want to evaluate students' overall mathematics proficiency at an aggregate level (e.g., province, region), a total score may be preferred. Similarly, a single metric of mathematics performance may be preferred for some program evaluation purposes (e.g., when using the scores as a way to understand the effects of various factors, such as gender or socioeconomic level). Relatedly, government officials without a strong background in early mathematics may have difficulty interpreting multiple pages of scores from individual subtests, each of which measures different foundational skills. Funders of large scale interventions may be unable to quickly grasp the implications of a report when multiple subscores are presented. For these and other uses, subscores do not provide the "at a glance" outcomes of which stakeholders have become accustomed from other mathematics assessments such as the TIMSS and PISA.

Because of these issues, users have sought alternate scoring methods for the EGMA, including reporting composite or total scores. Extending the scoring options for the EGMA may improve the accessibility and usability of the results for a variety of stakeholders. Composite scores may provide researchers with useful data to evaluate program or intervention effectiveness. In a recent example published by Piper et al. (2016), two composite scores were computed for the EGMA results: (1) subtests that assessed students' conceptual understanding and (2) those that assessed procedural fluency. These composite scores allowed the researchers to evaluate the effects of an intervention on two meaningful outcome variables.

Total scores may be useful when seeking to make group comparisons that support policy reforms or program evaluations. For example, in a cluster randomized controlled trial examining the impact of a distance education initiative on various indicators in Ghana, Johnston and Ksoll (2017) calculated a weighted total score for the EGMA (weighting was used to address the variability in the number of items per subtest). Similarly, analyzing policies in Ecuador, Cruz-Aguayo, Ibarraran, and Schady (2017) used total scores calculated from the EGMA to examine changes in students' mathematics performance within a school year based on teacher variables. However, while these test scoring methods may meet stakeholders' immediate needs, empirical evidence is needed to support the intended claim(s) that are associated with each scoring approach (Feinberg & Wainer, 2014). Different

scoring mechanisms impact the accuracy and interpretability of the results, which can have negative consequences.

The purpose of this paper is to examine three test scoring procedures for the EGMA and determine which approach(es) are psychometrically appropriate and useful. The three test scoring procedures examined are (1) total score (aggregate of correct responses across all items), (2) subscores, and (3) composite score (aggregate of subtest scores). We describe each scoring method in more detail and evaluate each method for reliability and distinctiveness of the results, and usefulness of the scores to relevant stakeholders. Although the principles discussed herein apply to scores derived using Item Response Theory (IRT) modeling, our discussion focuses on scores obtained using Classical Test Theory (CTT) approaches. The test scoring procedures are compared using data from an actual administration of the EGMA in Jordan. Conclusions and recommendations for test scoring procedures for the EGMA are made. Generalizations to other testing programs are proposed; however, because of the wide-spread use of the EGMA within the global mathematics education community, this manuscript is centrally focused on the EGMA.

Early Grade Mathematics Assessment

The EGMA is an orally and individually-administered assessment that measures young children's foundational mathematics knowledge. It is typically administered to students in Grades 1-3 and takes approximately 20 minutes to administer. The EGMA has been translated and adapted for use in many languages. The EGMA is composed of eight subtests. Each subtest includes 5-20 constructed-response items (i.e., students must provide a response on their own and are not given possible response options from which to choose). Table 1 details the subtests, time limits, and standard test scoring procedures as stated in the *Early Grade Mathematics Assessment (EGMA) Toolkit* published by RTI International (2014).

Three EGMA subtests are timed, and students have 60 seconds to generate responses. These subtests are typically scored as the *number of correct responses per minute*, and is calculated using the following equation:

$$NCPM = \frac{c \times 60}{t}$$

where: *NCPM* is the number correct per minute

c is the number of correct responses

t is the elapsed time in seconds taken by the student

This equation takes into consideration students who finish all items in less than 60 seconds. For example, if a student answers all 20 items correctly in 40 seconds, their score would be 30 correct items per minute, since they likely would have answered more items correctly if more items had been available.

The remaining five subtests are untimed and are scored as the total number of items correct. According to the administration procedures (RTI International, 2014), students must generate a response to each item within five seconds before the test administrator prompts the student to move to the next item. Additionally, these subtests have stopping rules, such that if a student answers four items in a row incorrectly, the test administrator stops the subtests and proceeds to the next subtest. The items on the EGMA are sequenced from least to most difficult (see RTI International [2014] for more details about item and subtest development). Therefore, if the stopping rule is applied, all of the remaining items are scored as incorrect, since the student likely would have responded incorrectly.

Table 1
Core EGMA Subtest Information (RTI International, 2013; Table modified from Perry, 2016)

Subtest	Number of Items	Task	Time Limit	Stopping Rule	Standard Test Scoring Procedure
Number Identification	20	Read numbers	60 seconds	None	Number correct per minute
Quantity Discrimination	10	Determine the larger of two numbers	No time limit	Stop the subtest if the child has four successive incorrect answers	Total number of items correct
Missing Number	10	Determine the missing number in a sequence of numbers	No time limit	Stop the subtest if the child has four successive incorrect answers	Total number of items correct
Addition – Level 1	20	Add two one-digit numbers	60 seconds	None	Number correct per minute
Subtraction – Level 1	20	Subtract two one-digit numbers	60 seconds	None	Number correct per minute
Addition – Level 2	5	Add a one-digit or two-digit number to a two-digit number	No time limit. This subtest is not administered to students who did not answer any items correctly on Level 1.	Stop the subtest if the child has four successive incorrect answers	Total number of items correct
Subtraction – Level 2	5	Subtract a one-digit or a two-digit number from a two-digit number	No time limit. This subtest is not administered to students who did not answer any items correctly on Level 1.	Stop the subtest if the child has four successive incorrect answers	Total number of items correct
Word Problems	6	Respond to a word problem read out loud	No time limit	Stop the subtest if the child has four successive incorrect answers	Total number of items correct

Scoring Procedures

Scoring of tests includes two distinct procedures. First, students' responses to items are scored following a set of guidelines to judge the correctness of the response. Second, the scored item responses are aggregated following another set of guidelines to arrive at one (or more) overall score for the test. The collection of scored item responses serve as evidence about students' levels of performance in the tested construct (Thissen & Wainer, 2001), and therefore, form the basis of test score uses and interpretations.

Consider a simplified example of the administration of a typical achievement test with multiple choice items. To score each item, a student's answer choice is compared to the correct answer. If the student selected the correct response from a given set of distractors, the response is coded as correct and the student is awarded a pre-specified number of points. To arrive at an overall test score using CTT, the number of correct responses or points can be summed to generate a raw score. The raw score can be converted to a ratio of number correct to total number of items (and reported as a ratio or percentage) or transformed to a standard score, which may be easier for some stakeholders to interpret. However generated, the overall test score is typically used to make judgements about the test taker's level of proficiency in the tested construct.

The selection of the item and test scoring procedures is a complex process that should align with the purpose of the test and support the intended uses and interpretations of the results (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; International Testing Commission [ITC], 2014). To some extent, item scoring procedures are influenced by the item format (i.e., selected response, constructed response). For example, constructed-response items ask students to construct their own response to an item and are often evaluated using a scoring rubric that details the response expectations associated with a specific score. Conversely, selected-response items ask students to select an answer from a set of possible responses, and can be scored following a dichotomous scoring rule that assigns value only to the correct response. Although these are typical practices, item scoring procedures may vary. Regardless of the item format, the item scoring procedures should support the intended uses and interpretations of the test scores.

Similarly, test scoring procedures need to provide test users with information that facilitates the intended uses and interpretations of the results. Test scoring begins with the specification of the scale on which scores will be reported, such as unweighted raw scores or model-derived scores such as those produced through Item Response Theory (IRT) modeling (Shaeffer et al., 2002). Test scores can be obtained for all items included on the test (e.g., total score), a subset of the items (e.g., subscores), or a collection of subsets of items (e.g., composite scores). The rationale and evidence supporting the alignment between these test scoring procedures and the purpose of the test should be documented (AERA, NCME, & APA, 2014). Furthermore, when more than the total score is reported, the reliability and distinctiveness of the subscores or composite scores should be provided to justify the appropriateness of the interpretations and uses. This paper focuses on evaluating possible scoring procedures for the EGMA.

Test Scoring Methods

Total Score

A total score is a summation of students' correct item responses across the overall test following the item-level scoring rules. Total scores are reported as one value. The reported value is intended to serve as an estimate of the student's overall level of proficiency in the tested construct. Students with similar total scores are considered to have similar levels of proficiency in the tested construct (Davidson et al., 2015).

The total score is calculated following specific scoring procedures that are outlined in the test specifications. The scoring procedures may specify differential weights to items or item types (e.g., constructed response) following a test blueprint. In some instances, the total score may be calculated from student's responses on subsections of a test that represent meaningful subcomponents of the construct but have too few items to allow for reliable estimates (Sinharay, Haberman, & Puhan, 2007).

For the EGMA, reporting a total score would represent a student's overall proficiency on early numeracy concepts. As noted in the introduction, stakeholders are frequently exposed to total scores. Policy makers may believe that an EGMA total score would be useful in evaluating the effectiveness of educational policies (similar to the example published by Cruz-Aguayo, Ibarraran, & Schady, 2017), providing a comprehensive measure of overall proficiency. Moreover, a single measure of mathematics proficiency may be useful for researchers examining the efficacy of an intervention on multiple outcome variables (as was reported by Johnston & Ksoll, 2017). Conversely, total scores may be less useful for policy makers interested in evaluating the effectiveness of curricular reforms or programs, or practitioners who want to evaluate the outcome of instructional practices or interventions on student learning.

Some concerns about reporting total scores have been raised in the literature. Davidson et al. (2015) point to possible unintended consequences of the assumption that test takers with similar scores have similar proficiency levels. Without considering the pattern of responses across the test, they argue that total scores may incorrectly cluster students on overall proficiency that might mask important differences across groups of students. For example, students scoring in the lower quartile of a test may have different patterns of errors that may point to important differences in their knowledge and skills on the tested construct. Reporting only the total score masks these differences.

Reporting total scores for the EGMA poses additional technical challenges. Namely, because each subtest includes a different number of items, simply summing the total number of correct responses would result in a differential weighting of some of the subtests. For example, there are 10 items on the Missing Number subtest and 5 items on the Word Problem subtest. If a student's responses are summed across these subtests, the student's performance on the Missing Number subtest would be given primacy to his or her performance on the Word Problem subtest.

Relatedly, as previously noted, the administration method varies across the subtests in that some are timed, and some are untimed. Certain analyses cannot be conducted when the timed and untimed items are combined together. For example, Cronbach's alpha values cannot be computed for the timed items because this coefficient does not take into consideration time, which is an important part of the scoring procedure. Confirmatory Factor Analysis can be used to estimate reliability of accuracy, where speed and accuracy are modeled jointly. However, this joint model would not be possible since accuracy (i.e., correct or not

correct) is measured at the item level but speed is measured at the subtest level. Reliability coefficients could be calculated for the timed subtests if both accuracy and speed were reported at the item level. This issue creates a ripple effect – the reliability of the total score of timed and untimed cannot be calculated, since the reliability cannot be calculated for the timed tests. These sources of variability in the composition and administration of the EGMA subtests may make reporting a total score technically complex and have implications for the interpretability of the summed scores. Possible alternatives to reporting total scores are to report subscores or composite scores.

Subscores

Subscores represent students' responses to items that assess specific and unique subcomponents of the overall construct (Sinharay, Puhan, & Haberman, 2011). Subscores are the most frequent method of reporting scores on EGMA assessments, though there are differences in whether or not the fluency measure (correct number per minute on timed tasks) is included (RTI International, 2014; Bridge International Academies, 2013). For a given testing situation, a student may receive multiple subscores, one for each subcomponent of the construct. For example, subscores for a comprehensive reading test might include vocabulary and reading comprehension. The reported scores are intended to provide more fine-grained information about students' level of proficiency in meaningful subcomponents of the construct. Provided that the subscores represent reliable and trustworthy data, the reported information can be used to make diagnostic inferences (Davidson et al., 2015).

For the EGMA, the subscores are associated with the individual subtests that comprise the full operational testing program. Because data are provided about students' performance on each concept that comprises early numeracy, these results may inform practitioners' interpretations about the effectiveness of instructional practices or interventions on student learning. These results may be directly applicable in classroom settings because they identify areas of strength and weakness that may guide teachers' instructional design and delivery making (Sinharay, Puhan, & Haberman, 2011).

Technical characteristics of subscores have been discussed in the literature. Subscores should provide useful information above that which is provided by the total score (Wedman & Lyren, 2015). Sinharay (2010) proposed that for subscores to have value they should be reliable and provide distinctive information. Reliability is necessary to provide stable estimates of students' performance from which decisions will be based (Feinberg & Wainer, 2014). Reliability may be compromised because of the small set of items often used to generate subscores (Stone, Ye, Zhu, & Lane, 2010). However, some of these limitations may be overcome if reporting data in aggregate form, such as reporting subscores for groups of students as opposed to individual students.

Subscores may be considered distinctive if they contribute unique information beyond the total score. Distinctiveness can be conceptualized as the degree of orthogonality between the subscores, and is often evaluated by examining the disattenuated correlation between subscores (Wedman & Lyren, 2015). That is, the smaller the correlation between the subscores, the greater the likelihood that the subtest is providing unique (or distinctive) information (Feinberg & Wainer, 2014). Sinharay (2010) analyzed results from a series of operational testing programs and simulation studies and found that the average disattenuated

correlations should be .80 or less to provide distinctive information.

Haberman (2008) proposed another approach to examining the usefulness of subscores, which combines the reliability coefficients and the disattenuated correlations of the subscores. Haberman's method (2008) examines the proportional reduction in mean squared error (PRMSE) values. PRMSE values range from 0 to 1, with larger values indicating more accurate measures of true scores with smaller mean squared errors. PRMSE values are calculated for the subscores ($PRMSE_s$) and then compared to the PRMSE values for the total or composite score ($PRMSE_x$). To add value, the $PRMSE_s$ must be greater than $PRMSE_x$. See Haberman (2008) for more information about this analytic method.

Research on the reliability and distinctiveness of subscores continues to emerge; however, notable concerns have been raised about the technical quality of subscores. Stone et al. (2010) identified a persistent problem with the reliability of subscores because of the limited number of items contributing to the scores. Similarly, Sinharay (2010) concluded that it is difficult to obtain reliable and distinctive subscores without at least 20 items. Moreover, if using subscores to evaluate changes in students' performance over time, additional methodological considerations must be taken into account when examining reliability (Sinharay & Haberman, 2015) that subsequently impact the ease of use in classroom settings.

Subscores are the standard mechanism by which student performance on the EGMA is reported (RTI International, 2014). Because the EGMA was designed to provide instructionally relevant information to score users, these data highlight strengths and areas for improvement that can be used to evaluate the effectiveness of instructional practices or interventions on student learning at the classroom level or for program evaluations. However, because of the limited number of items on each subtest, subscores are prone to be less reliable and more susceptible to floor (high proportion of minimum scores) and ceiling (high proportion of maximum scores) effects (RTI International, 2014). Of concern is the fact that increasing the number of items in all EGMA subtests to 20 would greatly increase the amount of time required to complete the assessment. This adds to costs and taxes students' attention over time.

In addition, providing multiple indicators of proficiency may compromise the interpretability of scores by policy makers or practitioners who are not familiar with the concepts that comprise early numeracy. A potential unintended consequence is the overgeneralization of subtest performance to curricular design decisions that results in narrowing the curriculum or teaching to the test. For example, the Missing Number subtest is intended to assess students' ability to interpret and reason about number patterns. If misinterpreted, results could be inappropriately used to instruct teachers to directly teach students to fill in a missing number from given sequences, as opposed to teaching the reasoning skills underlying the intention of the subtest. Some of these limitations have led policy makers and researchers to request composite scores.

Composite Scores

Composite scores represent aggregated student performance across meaningful components of the construct and, as such, are similar to subscores (Sinharay, Haberman, & Puhan, 2007). However, composite scores differ from subscores in that they may encompass more than one subtest, and/or may include items that represent different dimensions of the construct such as content classification (e.g., measurement, geometry) or process dimensions such as procedural knowledge and conceptual

understanding (Piper et al., 2016; Sinharay, Puhan, & Haberman, 2011; Stone et al., 2010). The hypothesized dimensions of the construct should be verified using appropriate analytic techniques such as factor analysis (Davidson et al., 2015). It follows that composite scores can be conceptualized as augmented subscores in which the subscores are weighted, either equally or differentially (Sinharay, 2010).

Composite scores may provide several advantages over subscores. Chiefly, composite scores typically include more items than subscores, which may improve score reliability. Also, because additional information contributes to the observed score, composite scores may increase the predictive utility of the outcome to a criterion (Davidson et al., 2015). Findings from operational testing programs and simulation studies suggest that composite scores add value more often than subscores as long as the disattenuated correlations were less than .95 (Sinharay, 2010).

For the EGMA, composite scores could be calculated by clustering subtests based on the assessed dimensions of early numeracy or the response processing requirements of the subtest. Because composite scores provide summary information that encompass meaningful dimensions of the construct, these data might help policy makers evaluate curricular reforms or programs by illustrating overall areas of strength or in need of improvement. These scores might be more interpretable than subscores, and may provide a better representation of students' proficiency in meaningful dimensions of early numeracy.

Composite scores can be based on specific subcomponents of the construct. For example, composite scores can be calculated for (1) Basic Number Concepts, which aggregates responses from the Number Identification, Quantity Discrimination, and Missing Number subtests, and (2) Operations and Applied Arithmetic, which aggregates responses from the Addition – Level 1, Addition – Level 2, Subtraction – Level 1, Subtraction – Level 2, and Word Problems subtests. These distinctions are based on research suggesting that early numeracy has a two-factor structure, with one factor focusing on basic number sense and number knowledge and the other factor focusing on problem solving and operations (Aunio, Niemivirta, Hautamäki, Van Luit, Shi, & Zhang, 2006; Jordan, Kaplan, Nabors Oláh, & Locuniak, 2006; Purpura & Lonigan, 2013).

Alternatively, composite scores can be based on response processing, and may include (1) untimed, which aggregates responses from the Quantity Discrimination, Missing Number, Word Problems, Addition – Level 2, and Subtraction – Level 2 subtests and (2) fluency of processing early numeracy concepts, which aggregates responses from the Number Identification, Addition – Level 1, and Subtraction – Level 1 subtests. Piper and colleagues (2016) created an index for procedural tasks (Number ID, Addition-Level 1, and Subtraction Level-1) and an index for conceptual tasks (all other untimed tasks) which aligned with the response processing described above. Other configurations of composite scores may be theoretically or substantively meaningful, depending on the outcomes of the program evaluation for which the EGMA is being used.

A persistent issue in computing composite scores is weighting of item sets or subtests. Differential weighting occurs either when item sets or subtests have different numbers of items or points to be aggregated, or when some item sets or subtests are more important or deserve greater emphasis in the composite score (Feldt,

2004). Differential weighting may also occur when using different item types. For example, Schaeffer et al. (2002) generated composite scores based on response type (i.e., selected response, constructed response) and investigated methodological solutions to address the differential weighting based on variability in the number of items for each response type.

These issues are pertinent to reporting composite scores for the EGMA. Because the item-level scoring approaches for the subtests on the EGMA vary, it is methodologically challenging to compute some composite scores, depending on the dimension to be aggregated. For example, as noted earlier, to calculate a composite score for Operations and Applied Arithmetic, students' responses could be aggregated for the Addition-Level 1, Addition-Level 2, Subtraction-Level 1, Subtraction-Level 2, and Word Problems subtests. The number of items, item-level scoring approach, and subtest scoring approach varies across these five subtests complicating the approach for computing a composite score.

To provide empirical evidence to evaluate the technical adequacy of these test scoring methods, data from an EGMA administration in Jordan in 2014 was used to examine the implications of different scoring procedures on the intended uses and interpretations of the test results.

Methods

Participants

We used an existing dataset obtained from an EGMA administration with 2,912 students in Jordan 2014. This dataset was used based on convenience. These data were particularly well suited for this study because the vast majority of children were appropriately-aged for the assessment and the language was stable across administrations. In addition, all of the subtests were administered.

For this study, data were removed for students who did not attempt at least one question on all EGMA subtests. Therefore, 60 cases were removed, leaving data from 2,852 students to be used in the analyses below. All students were in Grades 2-3. The average age was 8.33 years old (*SD* = 0.75). Additional information about the sample of students used for these analyses can be seen in Table 2. The EGMA was administered as part of an endline survey (meaning it was administered at the end of program implementation) to examine the impact of a literacy and mathematics intervention. RTI International managed the sampling procedures for the project. See Brombacher et al. (2014) for detailed information about sampling. A baseline survey (not used in this analysis) that examined students' foundational mathematics skills and associated Jordanian school-level variables served as the impetus for the intervention (Brombacher, 2015).

Table 2
Student characteristics for sample

Gender		Age in Years							School Location		Grade	
Female	Male	6	7	8	9	10	11	12	Urban	Rural	2nd	3rd
1,535	1,317	2	363	1,270	1,131	79	4	3	1,817	1,035	1,404	1,448

Instrument

All of the students took all eight EGMA subtests: Number Identification, Quantity Discrimination, Missing Number, Addition – Level 1, Addition – Level 2, Subtraction – Level 1, Subtraction – Level 2, Word Problems.

Administration procedures

A total of 56 test assessors administered the endline survey (Brombacher et al., 2014), and the majority of the assessors had previously administered the EGMA. The test assessors attended a 9-day training led by an RTI International employee on how to conduct the test administrations for the EGMA and Early Grade Reading Assessment (EGRA) endline surveys. Assessors practiced administering the EGMA with one another and practiced with students in area schools. Inter-rater reliability checks were conducted and a score of 0.90 or greater was required in order to assess students in the field.

The EGMA was administered using stimulus sheets that were seen by the students and tablets that assessors used to read the instructions for each subtest and to record students' answers. As previously noted, the EGMA is orally and individually administered. For the untimed subtests, test assessors were instructed to ask students to move to the next item if they had not responded in 5 seconds. Items that resulted in no response were left blank and were scored as incorrect.

Scoring

Items on the subtests were scored using each subtest's standard scoring procedure (see Table 1). The five untimed subtests were scored as the total number correct, and the three timed subtests were scored as the number correct per minute. Table 3 provides a summary of the subtest scores. As expected, there is greater variance in the scores for the timed subtests, since students could receive scores greater than the total number of items based on how much time remained when they completed the subtest (see previous section on EGMA scoring procedures). Additionally, the majority of the subtest scores are normally distributed with skewness and kurtosis values between (-1, 1). However, the Addition – Level 1 scores are highly leptokurtic (Kurtosis = 2.97).

Table 3
Summary of EGMA subtest scores

Subtest	Number of Items	Standard Scoring Procedure	*N*	Mean	SD	Maximum Score	Skewness	Kurtosis
NI	20	NCPM	2852	33.32	16.46	85.71	0.34	-0.41
QD	10	Total correct	2852	8.00	2.69	10	-1.42	1.07
MN	10	Total correct	2852	6.12	2.81	10	-0.35	-1.03
A1	20	NCPM	2852	12.61	5.29	50	0.39	2.97
S1	20	NCPM	2852	9.83	4.43	31.58	-0.06	0.99
A2	5	Total correct	2852	2.60	1.71	5	-0.02	-1.24
S2	5	Total correct	2852	1.75	1.68	5	0.61	-0.88
WP	6	Total correct	2852	3.58	1.82	6	0.34	-0.41

Analyses

Following recommendations proposed by researchers examining scoring procedures (c.f., Sinharay, 2010; Sinharay, Puhan, & Haberman, 2011; Stone, Ye, Zhu, & Lane, 2010; Wedman & Lyren, 2015), traditional reliability coefficients, disattenuated correlations, and proportional reduction in mean squared error (PRMSE) values were calculated to compare the reliabilities and distinctiveness of scores for the three test scoring methods for the EGMA (i.e., subscores, composite scores, total scores). The composite scores were based on the two-factor structure of early numeracy (Basic Number Concepts [BNC] and Operations and Applied Arithmetic [OAA]). As noted previously, composite scores can be created for different clusters of subtests. However, theoretical evidence about the nature of early numeracy supports this two-factor structure (c.f., Aunio, Niemivirta, Hautamäki, Van Luit, Shi, & Zhang, 2006; Jordan, Kaplan, Nabors Oláh, & Locuniak, 2006; Purpura & Lonigan, 2013).

For these analyses, we used only results from the untimed EGMA subtests. The scoring procedure for the timed subtests (i.e., number correct per minute) focuses on both accuracy and speed, and reliability coefficients cannot be calculated to consider both accuracy at the item-level and speed at the subtest-level. If data were collected on accuracy and speed at the item-level, reliability coefficients could be calculated using other methods. However, we were unable to apply a technically sound analytical approach to estimate reliability with the current parameters. As a result, the BNC composite score is calculated using results from the Quantity Discrimination and Missing Number subtests, but not for the Number Identification subtest (for clarity, we refer to this composite score as BNC-UT to note that it represents only the untimed subtests). Similarly, the OAA composite score only includes results from the Addition – Level 2, Subtraction – Level 2, and Word Problems subtests, but not the Addition – Level 1 and Subtraction – Level 1 (for clarity, we refer to this composite score as OAA-UT to note that it represents only the untimed subtests). Consequently, computing a total score is not possible; instead, we calculated an Overall Untimed Composite Score to include Quantity Discrimination, Missing Number, Addition – Level 2, Subtraction – Level 2, and Word Problems subtests.

Although traditional reliability coefficients have previously been calculated for these timed subtests, these estimates treat every item in the subtest, even those unreached, as incorrect/correct and do not consider the factor of time. Because this paper seeks to compare scoring procedures, we felt the need to ensure that all of the analyses conducted align with the subtests' scoring procedures and the interpretations made using those scores. Therefore, for the analyses described below, only data from the untimed tests was used. Implications for both untimed and timed tests are included in the discussion section.

Internal consistency reliability coefficients (i.e., Cronbach's alpha), were calculated in R (R Core Team, 2017) using the Psych package (Revelle, 2015) for each scoring procedure for the untimed EGMA subtests. We used guidelines proposed by Kline (2009) to evaluate the strength of the reliability coefficients. Kline suggests that coefficients for tests should be $\alpha > .7$, with $\alpha > .9$ indicating strong reliability and $.9 > \alpha > .7$ indicating moderately strong reliability. The strength of reliability depends on the use of the assessment. Low-stakes assessments should have moderately strong reliability coefficients, and high-stakes assessments should have strong reliability assessments.

In addition to being reliable, subscores (or composite scores) must also be distinct from other subscores (or composite scores). As

previously noted, distinctiveness can be evaluated by examining the disattenuated correlations (disattenuated from measurement error) between subscores. If subscores are too highly correlated, they do not add additional value or information beyond the total score. Therefore, in order for subscores (or composite scores) to be considered distinct, disattenuated correlations should be below 0.80 (Sinharay, 2010). Disattenuated correlations were calculated for the subscores in R (R Core Team, 2017) using the Psych package (Revelle, 2015).

Haberman's method (2008) was used to further examine the potential usefulness of the subscores. PRMSE values for the EGMA subscores ($PRMSE_s$) were compared to PRMSE values for the Overall Untimed Composite scores ($PRMSE_x$) to determine if the subscores add value over the Overall Untimed Composite score. In order to add value, $PRMSE_s$ must be greater than $PRMSE_x$, which indicates that the subscores reduce the mean squared error more than the Overall Untimed Composite score. $PRMSE_s$ and $PRMSE_x$ values were calculated in R (R Core Team, 2017) using the Subscore package (Dai, Wang, & Svetina, 2016).

Results

Internal consistency reliability coefficients, disattenuated correlations, and proportional reduction in mean squared error (PRMSE) values were calculated for the three test scoring methods for the EGMA (i.e., subscores, composite scores, total scores).

Reliability

Reliability coefficients are presented in Table 4. Internal consistency reliability for the subscores are moderately strong. The reliability coefficients for the composite scores (i.e., scores by construct) are also strong ($\alpha > .9$) to moderately strong, and the reliability of the Overall Composite score is strong. As expected, the more items included in a score, the higher the reliability of the score. See Table 4 for the results.

Distinctiveness of Scores

Disattenuated correlations were calculated for subscores and composite scores. All of the disattenuated correlations for the subscores are above the diagonal of reliability coefficients and are less than 0.80 (see Table 5), except for the disattenuated correlation between Addition – Level 2 and Subtraction – Level 2, which was 0.88. These findings provide evidence that the subscores are distinct and provide additional information, with the exception of Addition – Level 2 and Subtraction – Level 2.

Table 4

Cronbach's alpha coefficients by scoring procedure for untimed EGMA subtests

Subtest	Cronbach's Alpha		
	By Subscore	By Untimed Composite	By Overall Untimed Composite Score
Quantity Discrimination	0.88	0.91	0.94
Missing Number	0.86		
Addition – Level 2	0.78	0.88	
Subtraction – Level 2	0.79		
Word Problems	0.74		

Table 5
Reliability coefficients (on diagonal), correlations (below diagonal), and disattenuated correlations (above diagonal) for subscores

Subtest	QD	MN	A2	S2	WP
Quantity Discrimination (QD)	**0.88**	0.77	0.61	0.52	0.64
Missing Number (MN)	0.67	**0.86**	0.74	0.70	0.78
Addition – Level 2 (A2)	0.50	0.60	**0.78**	0.88	0.72
Subtraction – Level 2 (S2)	0.44	0.57	0.70	**0.79**	0.72
Word Problems (WP)	0.52	0.62	0.55	0.55	**0.74**

Next, the disattenuated correlations were calculated for the composite scores based the two-factor structure of early numeracy (BNC-UT and OAA-UT) (see Table 6). The disattenuated correlation between the BNC-UT and OAA-UT scores is 0.77, indicating that the composite scores based on construct are distinct.

Table 6
Reliability coefficients (on diagonal), correlations (below diagonal), and disattenuated correlations (above diagonal) for composite scores based on construct

	OAA-UT	BNC-UT
Operations and Applied Arithmetic (OAA)	**0.91**	0.77
Basic Number Concepts (BNC)	0.70	**0.88**

Haberman's Method (2008)

To implement Haberman's method (2008), PRMSE values were calculated for the subscores (see Table 7). For each of the subscores, the $PRMSE_s$ values are greater than the $PRMSE_x$ values, indicating that the subscores add value beyond that of just the Overall Untimed Composite score. The subscores, compared to the Overall Untimed Composite score, provide more accurate estimates of the true score.

Next, $PRMSE_s$ values were calculated for the BNC-UT and OAA-UT composite scores (see Table 8). For each of the composite scores, the $PRMSE_s$ values are greater than the $PRMSE_x$ values, indicating that the composite scores add value over the Overall Untimed Composite score. The BNC-UT and OAA-UT composite scores, compared to the Overall Untimed Composite score, provide more accurate estimates of the true score.

Table 7
Proportional reduction in mean squared error (PRMSE) for subscores

Subtest	$PRMSE_s$	$PRMSE_x$
Quantity Discrimination	0.88	0.69
Missing Number	0.86	0.82
Addition – Level 2	0.73	0.72
Subtraction – Level 2	0.78	0.66
Word Problems	0.79	0.71

Table 8
Proportional reduction in mean squared error (PRMSE) for composite scores

Construct	Subtests	$PRMSE_s$	$PRMSE_x$
Basic Number Concepts-Untimed	Quantity Discrimination Missing Number	0.91	0.85
Operations and Applied Arithmetic-Untimed	Addition – Level 2 Subtraction – Level 2 Word Problems	0.88	0.82

Discussion

The purpose of this manuscript was to examine three test scoring approaches for the EGMA to address a stated need in the field to provide various stakeholders with actionable and interpretable results. The criteria on which the test scoring approaches were evaluated included the psychometric properties and usefulness of the results to stakeholders. Each test scoring approach was evaluated against these criteria, and implications for the validity of the intended uses and interpretations are considered.

Evaluation of the Psychometric Properties of Three Test Scoring Approaches

As previously noted, responses from the timed EGMA subtests cannot be combined with responses from the untimed EGMA subtests because such procedures conflict with generally accepted statistical tenets. Because of this technical limitation, generalizations about the total scores and composite scores are based on aggregating results from the untimed subtests including Quantity Discrimination, Missing Number, Addition – Level 2, Subtraction – Level 2, and Word Problems. Taking this constraint into account, it is not possible to generate a total score for the operational EGMA, or composite scores that are fully representational of the subcomponents of the construct of early numeracy. As such, the discussion about the psychometric properties will focus on subscores and three composite scores with untimed subtests (OAA-UT, BNC-UT, and Overall Untimed Composite).

Three psychometric properties were examined: internal consistency reliability, distinctiveness of the scores, and the additional information provided by the scores. First, the internal consistency reliability coefficients were examined for each subscore, the OAA-UT and BNC-UT composite scores, and the Overall Untimed Composite scores. All reliability coefficients were within acceptable bounds. Second, when examining the distinctiveness of the subscores, all subscores are distinct, with the exception of Addition – Level 2 and Subtraction – Level 2. It is possible that responses from these subtests can be combined to improve the distinctiveness of these subscores. Moreover, the OAA-UT and BNC-UT composite scores are distinct. Third, and finally, to examine the value of the information provided by the subscores and the composite scores, the proportion reduction in mean squared error (PRMSE) was

examined for the subscores and composite scores as compared to the Overall Untimed Composite score. Results indicate that the subscores and OAA-UT and BNC-UT composite scores add value beyond the Overall Untimed Composite score and provide more accurate estimates of the true score. In summary, the available evidence supports the psychometric properties of the subscores, and the OAA-UT and BNC-UT composite scores. Because the subscores and OAA-UT and BNC-UT composite scores provide more accurate estimates of the true score than the Overall Untimed Composite score, the use of the Overall Untimed Composite score is not supported with the psychometric evidence obtained in these analyses.

Evaluation of the Usefulness of Test Scoring Approaches to Stakeholders

An important consideration for this manuscript was the usability of the results by various stakeholders. As previously noted, stakeholders use the EGMA results for different purposes and, as such, may seek different mechanisms of aggregating students' responses. Concerns regarding the interpretability of the subscores have emerged from the field, specifically focused on the length of score reports and presentation of results, as well as the difficulty non-experts face when deciphering the information. In this section, we examine the usefulness of the test scoring approaches for guiding reforms in policies and practices, and informing intervention design and evaluation.

To aid in determining if the results are useful for making these decisions, we selected specific cases of students to illustrate the implications of reporting total scores, composite scores, and subscores on subsequent decisions. These cases represent actual children within the dataset used for these analyses.

Because stakeholders may aggregate subtest scores without knowing the limitations of the psychometric properties of the scoring approaches, we examine a range of test scoring approaches. For the total score, we examine (1) Total Score that includes all eight subtests, and (2) Overall Untimed Composite Score that includes only the untimed subtests. For the composite scores, we examine (1) Comprehensive Composite Scores that include all subtests that contribute to the associated subcomponents of the construct (OAA and BNC), and (2) Untimed Composite Scores that include only the untimed subtests that contribute to the subcomponents of the construct (OAA-UT and BNC-UT). All subscores for the subtests are considered. It is important to note that the scores for the timed subtests (e.g., NI, A1, S1) may exceed the total number of items. Based on the formula presented earlier, this situation occurs when a student responds to all of the items in less than 60 seconds.

From the operational data, we selected eight cases that are clustered into three groups based on the Total Score. Cases in Group 1 have high Total Scores, cases in Group 2 have moderate Total Scores, and cases in Group 3 have comparatively lower Total Scores. These data are presented in Table 9. The mean and standard deviation for the Total Scores are M=77.81 and SD=30.99. The distribution of Total Scores can be seen in Figure 1.

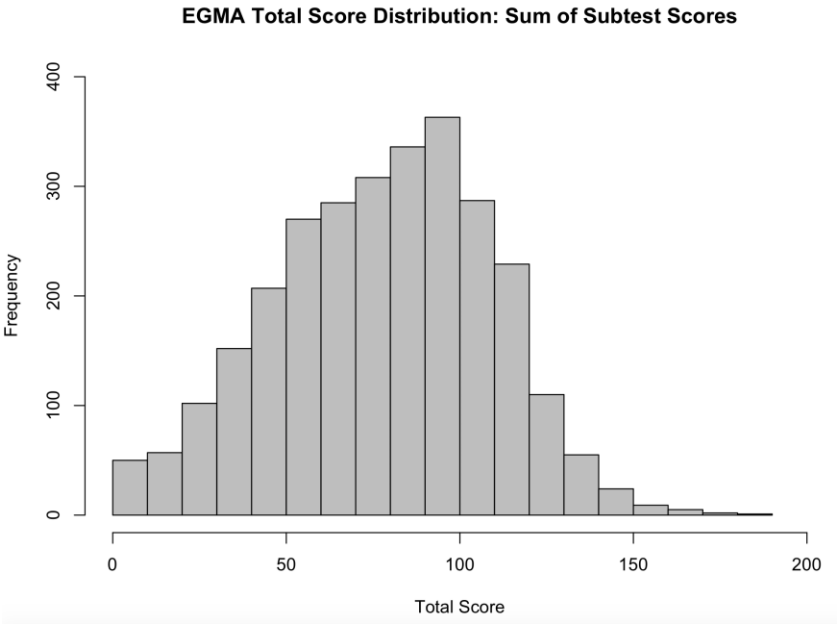

Figure 1. EGMA Total Score Distribution.

Interpretations Based on Total Scores

Because of the similarities in the Total Scores for students in these groups, stakeholders may conclude that the students in each group have similar proficiency levels in early numeracy. As noted earlier, policy makers may seek to use total scores to help evaluate the effectiveness of educational policies or curricular reforms, and researchers or practitioners may look to total scores to evaluate the outcomes of instructional practices or programs. However, by examining the total scores, important differences in students' levels of proficiency may be masked. For example, further examination of the Comprehensive Composite Scores indicate that differences may exist in the students' levels of proficiency in OAA and BNC. Notably for Group 1, Student A appears to have stronger BNC than OAA; whereas, Student B appears to have similar levels of proficiency in both subcomponents of the construct. Similar observations can be made for cases in Groups 2 (e.g., Students C and D, respectively) and 3 (Students G and H, respectively). As such, using the Total Scores to make decisions about the effectiveness of policies, curricular reforms, and instructional programs may lead to inaccurate conclusions.

These observations may be explained by differential weighting of the subtests that results from the variability in the number of items per subtest and the administration procedures leading to differences in the score units (e.g., raw score for untimed subtests, rate of correct

responses for timed subtests). For subtests with greater numbers of items, their proportion of contribution to the total score is increased. Thus, the skills and knowledge that are assessed on these subtests receive greater emphasis in the calculation of the total score. Similarly, the timed subtests have considerably higher score ranges because they are reported as a rate.

Controlling for the variability in the administration procedures, we can examine the Total Untimed Composite Scores. Omitting the timed subtests when calculating a total score leads to different groupings of students based on overall proficiency levels. Student B (shaded in dark grey) stands out as having the highest level of proficiency, followed by Students A and D (shaded in medium grey), then Students C and E (shaded in light grey), and Students F-H remain unshaded with the lowest Total Untimed Composite Scores. However, the aggregated score continues to mask some differences in students' levels of proficiency that are apparent when examining the Untimed Composite Scores (BNC-UT and OAA-UT). Although Students A and D have similar patterns of correct responses on BNC-UT and OAA-UT, Students C and E appear to have different levels of proficiency in BNC-UT and OAA-UT that are masked by similar Total Untimed Composite Scores. Parallel observations are noted for Students F and H. Comparable to the cautions noted when examining the total scores, examining the Total Untimed Composite scores may lead to inaccurate conclusions about the effectiveness of policies, curricular reforms, and instructional programs.

A possible solution that could address the differential weighting of subtests is to calculate a ratio of correct to incorrect responses for each subtest and then aggregate these ratios. However, as noted at the beginning of this manuscript, the impetus for this research was to address a need in the field for more interpretable reports. Creating and aggregating score ratios may not support this aim.

Interpretations Based on Composite Scores

Some stakeholders have called for composite scores to increase the interpretability, and thus usefulness, of the EGMA results for making decisions. Examining the Comprehensive Composite Scores (BNC and OAA) presented in Table 9, it is evident that additional information is provided about specific strengths and areas for growth in students' understanding of early numeracy concepts. This information may provide useful insights into aspects of policies, reforms, or programs that are or are not supporting students' learning of these important dimensions of early numeracy. However, just as was observed when analyzing the usefulness of the Total Score, these composite scores are heavily influenced by the extreme range of scores possible in the timed subtests, which differentially weights the scores in favor of these subtests. For example, Student A scored 70.59 on Number Identification; his or her scores on the remaining seven subtests combine to total 60. As such, the usefulness and interpretability of these scores may be compromised.

Table 9
Example student data from EGMA administration by scoring procedure

Group	Student	Subscores								Untimed Composite Scores		Total Untimed Composite Score	Comprehensive Composite – All Subtests		Total Score – All
		NI	A1	S1	QD	MN	A2	S2	WP	BNC-UT	OAA-UT		BNC	OAA	
Max Score		NA	NA	NA	10	10	5	5	6	20	16	36	NA	NA	NA
1	A	70.59	13	10	10	9	2	3	3	19	8	27	89.59	31	120.59
	B	39.31	32.43	15	10	10	5	5	6	20	16	36	59.31	63.43	122.74
2	C	46	4	5	10	7	1	0	3	17	4	21	63	13	76
	D	18	17.87	10	10	10	3	3	4	20	10	30	38	37.87	75.87
	E	33.53	12	11	6	5	2	2	4	11	8	19	44.53	31	75.73
3	F	6	11	11	1	3	2	1	4	4	7	11	10	29	39
	G	26.67	3	2	3	2	0	0	1	5	1	6	31.67	6	37.67
	H	11	9	4	9	0	2	0	2	9	4	13	20	17	37

Note: Number Identification (NI), Addition – Level 1 (A1), Subtraction – Level 1 (S1), Quantity Discrimination (QD), Missing Number (MN), Addition – Level 2 (A2), Subtraction – Level 2 (S2), Word Problems (WP). The Composite - Untimed scores include QD + MN for BNC and A2 + S2 + WP for OAA. The Composite - with Timed scores includes all subtests: NI + QD + MN for BNC and A1 + S2 + A2 + S2 + WP for OAA.

Again, to control for the variability in administration procedures, we examine the Untimed Composite Scores (BNC-UT and OAA-UT). Although the range of values in these scores is constrained, two additional problems are evident. First, within the OAA-UT composite score, the subtests do not have the same number of items, such that scores from the Word Problems subtest account for a larger proportion of the score than Addition-Level 2 and Subtraction-Level 2. The second and more significant issues is that these composite scores under-representation of the subcomponents of the construct. Because BNC-UT and OAA-UT are based on a truncated set of subtests, they are not inclusive of the range of knowledge and skills that define the two-factor structure of early numeracy. Thus, the meaningfulness and trustworthiness of the Untimed Composite Scores for guiding decisions about students' knowledge, skills, and ability in early numeracy may be compromised.

Conclusions

Several limitations impact the generalizability of these results. First, the composite scores used in these analyses were based on a subset of the EGMA subtests that most closely aligned with the research on the two-factor model of early mathematics. However, the composite scores could be created using different clusters of subtests. Changing the subtests would alter the composite scores, and may impact the outcomes of this study. Second, this study was conducted using a convenience sample from one country. This sample may have unique characteristics that do not generalize. Conducting these analyses with data from other countries would strengthen the generalizability of the findings.

In sum, based on the psychometric properties and usefulness of scores derived from three test scoring procedures, the evidence points to the need to continue reporting and using the subscores for the EGMA subtests when disseminating results. Although psychometrically adequate, composite scores based on the untimed subtests may distort the interpretations of students' levels of proficiency in early numeracy because they are based on a limited set of subtests. Subscores on the EGMA subtests provide detailed information about students' levels of proficiency on each concept that comprises early numeracy. These results can be used to evaluate the effectiveness of policies, curricular reforms, and/or instruction and intervention design.

References

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

Aunio, P., Niemivirta, M., Hautamäki, J., Van Luit, J. E. H., Shi, J., & Zhang, M. (2006). Young children's number sense in China and Finland. *Scandinavian Journal of Educational Research*, *50*(5), 483-502.

Bridge International Academies. (2014). *The Bridge Effect: Comparison of Bridge Pupils to Peers at Nearby Schools*. Nairobi, Kenya: Bridge International Academies. Retrieved from http://www.bridgeinternationalacademies.com on 6/11/2018.

Brombacher, A., Stern, J., Nordstrum, L., Cummiskey, C., & Mulcahy-Duhn, A. (2014). Education data for decision making (EdData II): National early grade literacy and numeracy survey – Jordan. Research Triangle Park, NC: RTI International.

Brombacher, A. (2015). National intervention research activity for early grade mathematics in Jordan. In X. Sun, B, Kaur, & J. Novotná (Eds.) Conference Proceedings of the Twenty-third ICMI Study: Primary Mathematics Study on Whole Numbers.

Brombacher, A., Bulat, J., King, S., Kochetkova, K., and Nordstrum, L. (2015). National Assessment Survey of Learning Achievement at Grade 2: Results for early grade reading and mathematics in Zambia. Research Triangle Park, NC: RTI International.

Cruz-Aguayo, Y., Ibarraran, P., & Schady, N. (2017). *Do Tests Applied to Teachers Predict their Effectiveness* (IDB Working Paper Series No IDB-WP-821). Washington, DC: Inter-American Development Bank.

Dai, S., Wang, X., & Svetina, D. (2016). *Subscore: Computing subscores in Classical Test Theory and Item Response Theory*. R package. Bloomington, Indiana: Indiana University.

Davidson, M. L., Davenport, E. C., Chang, Y-F., Vue, K., & Su, S. (2015). Criterion-related validity: Assessing the value of subscores. *Journal of Educational Measurement, 52*(3), 263-279.

Feinberg, R. A., & Wainer, H. (2014). When can we improve subscores by making them shorter?: The case against subscores with overlapping items. *Educational Measurement: Issues and Practice, 33*(3), 47-54.

Feldt, L. S. (2004). Estimating the reliability of a test battery composite or a test score based on weighted item scoring. *Measurement and Evaluation in*

Counseling and Development, 37, 184-190.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

International Test Commission. (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*(3), 195-217.

Johnston, J., & Ksoll, C. (2017). *Effectiveness of Interactive Satellite-Transmitted Instruction: Experimental Evidence from Ghanaian Primary Schools* (CEPA Working Paper No. 17-08). Palo Alto, CA: Stanford Center for Education Policy Analysis.

Jordan, N. C., Kaplan, D., Nabors Oláh, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development*, *77*(1), 153-175.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London: Routledge.

Perry, L. E. (2016). *Validating interpretations about student performance from the Early Grade Mathematics Assessment relational reasoning and spatial reasoning subtasks* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Global. (Order No. 10164141).

Piper, B., & Mugenda, A. (2014). *The Primary Math and Reading (PRIMR) Initiative: Endline Impact Evaluation*. Research Triangle Park, NC: RTI International.

Piper, B., Ralaingita, W., Akach, L., & King, S. (2016). Improving procedural and conceptual mathematics outcomes: Evidence from a randomised controlled trial in Kenya. *Journal of Developmental Effectiveness* Published online March 21, 2016. doi: 10.1080/19439342.2016.11

Platas, L.M., Ketterlin-Geller, L.R., & Sitabkhan, Y. (2016). Using an assessment of early mathematical knowledge and skills to inform policy and practice: Examples from the early grade mathematics assessment. *International Journal of Education in Mathematics, Science and Technology, 4*(3), 163-173. DOI:10.18404/ijemst.20881

Purpura, D. J., & Lonigan, C. J. (2013). Informal numeracy skills: The structure and relations among numbering, relations, and arithmetic operations in preschool. *American Educational Research Journal*, *50*(1), 178-209.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revelle, W. (2015). *psych: Procedures for personality and psychological research*. R package. Evanston, Illinois: Northwestern University.

RTI International. (2014). *Early Grade Mathematics Assessment (EGMA) Toolkit*. Research Triangle, NC: RTI International

Schaeffer, G. A., Henderson-Montero, D., Julian, M., & Bene, N. H. (2002). A comparison of three scoring methods for tests with selected-response and constructed-response items. *Educational Assessment, 8*(4), 317-340.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150-174.

Sinharay, S., & Haberman, S. J. (2015). Comments on “A Note on Subscores” by Samuel A. Livingston. *Educational*

Measurement: Issues and Practice, 34(2), 6-7.

Sinharay, S. Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21-28.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29-40.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*(1), 63-86.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Torrente, C., Aber, J.L., & Shivshaker, A. (2011). *Opportunities for Equitable Access to Quality Basic Education (OPEQ): Baseline Report: Results from the Early Grade Reading Assessment, the Early Grade Math Assessment, and children's demographic data in Katanga Province, DRC*. New York: New York University.

Wedman, J., & Lyren, P. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research, & Evaluation*, *20*(21).

About the Author(s)

Leanne Ketterlin Geller, PhD, is the Texas Instruments Endowed Chair in Education and professor in the Simmons School of Education and Human Development at Southern Methodist University. Her research focuses on supporting student achievement in mathematics through developing technically rigorous formative assessment procedures and effective classroom practices. Her work emphasizes valid decision-making systems for students with diverse needs.

Lindsey Perry, PhD, is a Research Assistant Professor at Southern Methodist University. Her current research interests focus on investigating children's spatial and relational reasoning abilities, developing mathematics assessments for young children, and training educators on how to use data from assessments to make instructional decisions.

Linda M. Platas, PhD, is the associate chair of the Child and Adolescent Development Department at SF State University. She has participated in the development of child assessment instruments including the Early Grades Math Assessment (EGMA) and the Measuring Early Learning Quality and Outcomes (MELQO) and served as an expert in mathematics and literacy development on many technical and policy groups. She is a member of the Development and Research in Early Math Education (DREME) Network.

Dr. Yasmin Sitabkhan, PhD, is a Senior Early Childhood Education Researcher and Advisor in RTI's International Education Division. In her current role at RTI, Dr. Sitabkhan provides technical support to projects in low- and middle-income countries in early mathematics. Her research interests focus on children's development of early mathematical concepts and instructional strategies to support learning in low- and middle-income contexts. Dr. Sitabkhan has a Ph.D. in Education from the University of California, Berkeley.