Invitational Research Symposium on
Through-Course
Summative Assessments

GENERALIZABILITY AND RELIABILITY: APPROACHES FOR THROUGH-COURSE ASSESSMENTS

Michael J. Kolen

The University of Iowa

March 2011

Center for K – 12 Assessment
& Performance Management at ETS

Executive Summary

Systems to assess the Common Core State Standards (Council of Chief State School Officers, & National Governors Association Center, 2010) have been proposed by the Partnership for Assessment of Readiness for College and Careers (PARCC, 2010) and the SMARTER Balanced Assessment Consortium (SBAC, 2010). The assessments proposed by these consortia involve assessing students at various times during the academic year and reporting a variety of different types of test score information that includes the following:

- Scores that are weighted composites over different test components
- Scores based on a single task or on a small number of tasks that are constructed-response tasks scored by humans and/or computer
- Raw and/or scale scores on various components of the assessments
- Achievement level classifications
- Subscores and/or content cluster scores
- Scores for aggregates (e.g., classrooms, schools, districts, states)
- Growth indices for individuals and aggregates

The large number of scores and score types, along with the complexity of many of the assessment tasks, suggest that the assessment of reliability and the use of reliability information in planning for these assessments will be challenging.

Recommendations

The abbreviated recommendations are as follows.

Recommendation 1

Conduct pilot studies during development of the PARCC and SBAC assessments that allow for the estimation of reliability using different numbers of tasks and raters. The results from these pilot studies can be used to refine the assessment tasks, rating procedures, and

assessment design so that scores on the constructed-response components of the assessments are of adequate reliability.

Recommendation 2

To the extent possible, develop constructed-response tasks that consist of a number of separately scored components that could lead to more reliable scores than if one holistic score is used with each constructed-response task.

Recommendation 3

(a) Estimate reliability of scores based on human judgment and use these to represent reliability of scores based on automated scoring. (b) Conduct research on procedures for assessing the reliability of scores that are based on automated scoring systems.

Recommendation 4

Use psychometric methods that do *not* assume that student proficiency is constant over various times of administration. Instead, estimate reliability for each component separately and use psychometric procedures that are designed to assess reliability for composite scores.

Recommendation 5

Use psychometric methods that do *not* assume that student proficiency is the same over different task types. Instead, estimate reliability for each component separately and use psychometric procedures that are designed to assess reliability for composite scores.

Recommendation 6

In developing the weights for each component of weighted composites, balance the practical need to give substantial weight to the constructed-response task components and the psychometric need to have weighted composite scores that are adequately reliable for their intended purposes.

Recommendation 7

(a) During development of the assessments, conduct pilot studies to estimate the reliability of the scores and modify the assessments, where needed, to achieve adequate score reliability. (b) Assess the reliability of each of the different types of scores for the assessments that are administered operationally.

Recommendation 8

(a) During development of the assessments, conduct pilot studies to estimate the reliability for each subgroup (including English language learners and students with various disabilities) and modify the assessments where needed to achieve adequate reliability for all students. (b) Assess score reliability for each subgroup for the assessments that are administered operationally.

Generalizability and Reliability: Approaches for Through-Course Assessments

Michael J. Kolen

The University of Iowa

Introduction

Systems to assess the Common Core State Standards (Council of Chief State School Officers, & National Governors Association Center, 2010) have been proposed by the Partnership for Assessment of Readiness for College and Careers (PARCC, 2010) and the SMARTER Balanced Assessment Consortium (SBAC, 2010). The assessments proposed by these consortia involve assessing students at various times during the academic year and reporting a variety of different types of test score information that includes the following:

- Scores that are weighted composites over different test components
- Scores based on a single task or on a small number of tasks that are constructed-response tasks scored by humans and/or computer
- Raw and/or scale scores on various components of the assessments
- Achievement level classifications
- Subscores and/or content cluster scores
- Scores for aggregates (e.g., classrooms, schools, districts, states)
- Growth indices for individuals and aggregates

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, NCME], 1999), Standard 2.1, "for each total score,

subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors or test information functions should be reported" (p. 31). Thus, reliability information for each of the scores proposed by the consortia will need to be provided. In addition, reliability information for each of the scores will need to be collected in pilot studies conducted prior to operational administrations so that scores on the assessments that are implemented will have adequate reliability for the intended purposes of the assessments. The large number of scores and score types, along with the complexity of many of the assessment tasks, suggest that the assessment of reliability and the use of reliability information in planning for these assessments will be challenging.

The purpose of this paper is to discuss these challenges and suggest solutions for estimating generalizability and reliability of the scores to be reported for the PARCC (2010) and SBAC (2010) assessments. The paper begins by describing concepts related to generalizability and reliability that are pertinent to these assessments. The concepts are then applied to the assessment systems proposed in the PARCC and SBAC and summarized in Educational Testing Service (ETS; 2010). The paper concludes with a summary of challenges and a set of recommendations.

Generalizability and Reliability Concepts

Generalizability and *reliability* are terms used to denote the precision or consistency of scores. The term *measurement error* is used to indicate the amount of imprecision in test scores. As measurement error increases, generalizability and reliability decrease.

Measurement precision is typically assessed using a psychometric model of one of the following types: *generalizability theory* (Brennan, 2001; Haertel, 2006), *classical test theory* (Haertel), or *item response theory* (IRT; Yen & Fitzpatrick, 2006). Generalizability theory was developed to comprehensively assess the influences of different sources of measurement error on score precision. Classical test theory is unable to distinguish among different sources of measurement error (Haertel). IRT also does not typically distinguish among different sources of

measurement error (Bock, Brennan, & Muraki, 2002; Brennan, 2006, p. 7). In this paper, the conceptual framework for considering precision of scores is based on the comprehensiveness of generalizability theory. However, the term *reliability* will be used as a general term that refers to the precision of scores from generalizability theory, classical test theory, or IRT perspectives.

Replications

To define the concept of reliability, it is necessary to consider *replications* of a *measurement procedure* (Brennan, 2006). The term *measurement procedure* is a general term used to denote a test or assessment for which a score is produced. The term *replication* refers to repeated applications of a measurement procedure. What constitutes a replication must be clearly specified to define the reliability of a measurement procedure.

For the purposes of the present paper, replications based on two different characteristics are considered: tasks and raters. *Tasks* refer to the test items that the examinee responds to and for which a score is provided. *Raters* refer to the individuals who score the examinee responses to the tasks when constructed-response tasks are used. In addition to tasks and raters, it also is important to specify the characteristics of the examinees that will be assessed by the measurement procedure.

To consider the reliability of a measurement procedure, it is necessary to specify what constitutes a replication for each of these characteristics and to specify what constitutes the standardized conditions for administering the assessment. When considering tasks, it is necessary to specify what constitutes an acceptable task or set of tasks along with the scoring rubric for each task. In the test construction process, the acceptable set of tasks is often operationalized through detailed content and statistical specifications. These specifications are then used to define the characteristics of a test form for a fixed-form test and a task pool for a computer-adaptive test. Any test form or task pool that meets the specifications is considered to be a replication of a set of tasks for the measurement procedure.

When raters are used to score examinee responses to constructed-response tasks, it is important to have detailed specifications for rater characteristics such as education, experience, and training. After these characteristics are clearly specified, any representative set of raters scoring the constructed response to a particular task with a well-defined scoring rubric is considered to be a replication of the rater characteristic for the measurement procedure.

Based on the administration of an assessment to an examinee, the score the examinee receives depends on the tasks that were administered and the raters used to score the constructed-response tasks. We could conceive of administering different tasks to this examinee, where different raters score the responses. If this measurement procedure were, hypothetically, repeated many times, the average score would be the *true score* in classical test theory and IRT or the *universe score* in generalizability theory. The consistency of the scores would be an indication of measurement precision, and the variability of the scores would be an indication of the amount of measurement error.

Ideally, reliability would be estimated by administering replications of the measurement procedure to each examinee. However, it is typically not feasible to assess an examinee more than one time. Instead, psychometric models are used, along with a set of assumptions required to use the model, to estimate reliability using the responses of examinees to tasks from a single administration of an assessment. The key to using any of these models is the ability to use parts of the test to define a replication. For example, in classical test theory, split-halves reliability might be estimated by dividing a test into two parts that are considered to be equivalent measures of the same construct. These two parts would be considered as replications of the task characteristic. Assumptions of classical test theory would be used to project reliability for the full-length test. There is a variety of other ways to assess reliability from a single administration using classical test theory and generalizability theory (Brennan, 2001; Haertel, 2006) and using IRT (Yen & Fitzpatrick, 2006). Each of the procedures requires

that replication be explicitly defined, which is the key concept that allows psychometric model–based procedures to be used to estimate reliability.

When constructed-response tasks are scored by raters, each response to a task is often scored by more than one rater. *Rater agreement* refers to the stability of scores across raters. When considering rater agreement, only raters vary, and not tasks. Thus, even though rater agreement is sometimes referred to as *rater reliability*, this terminology is a misnomer, because there is no variation of tasks.

Standard Error of Measurement

The *conditional standard error of measurement* for an individual, defined as the standard deviation of scores for an individual over replications, is often used as an index of the amount of variability in scores over replications. *Average standard errors of measurement* are indices of the conditional standard error of measurement averaged over examinees in the population. Standard errors of measurement are reported in the units of the scores of the assessment. Generalizability theory, classical test theory, and IRT all have procedures for estimating standard errors of measurement, though the estimation procedures differ and definitions of measurement errors differ across theories. In addition, the concept of *test information,* which is the reciprocal of measurement error variance, is used in IRT (Yen & Fitzpatrick, 2006).

Coefficients

Generalizability and reliability coefficients are indices of the precision of scores over replications for a population of examinees. Such coefficients range from 0 to 1 and are conceptualized as the proportion of overall variability in scores that is attributable to the true or universe scores. If the generalizability or reliability coefficient is 1, there is no measurement error. If the generalizability or reliability coefficient is 0, the measurements contain only

measurement error. Generalizability theory, classical test theory, and IRT all have procedures for estimating coefficients.

Indices for Aggregates (e.g., Classrooms, Schools, Districts)

The discussion so far has been concerned with scores of individuals. Generalizability theory and classical test theory have associate procedures for the estimation of the standard error of measurement and reliability of average scores for aggregates, such as at the level of the classroom, school, district, or state. Consideration of reliability of scores for aggregates is sometimes used when test scores are used for accountability purposes.

Decision Consistency for Proficiency Levels

In an accountability context, the focus of score reporting often is on proficiency levels, such as *basic, proficient*, and *advanced*. When there is a small number of levels, *decision consistency coefficients* are used. These coefficients are intended to reflect the proportion of examinees that are classified in the same proficiency level on two replications of an assessment. Generalizability theory, classical test theory, and IRT all have procedures for estimating decision consistency.

Composite Scores

The score used for accountability purposes is often a composite score over a set of assessments. Measurement error typically is assumed to be independent from one assessment to another. By assuming such independence, conditional standard errors of measurement for the composite can be calculated by summing squared conditional standard errors of measurement (referred to as *conditional error variances*) over the scores used to form the composite and then taking the square root of the sum. A similar process can be used to find the average standard error of measurement. Reliability coefficients can be calculated by subtracting the ratio of average error variance for the composite to total variance for the composite from 1. Thus, when composite scores are used, the reliability coefficient for the

composite can be calculated using information about the reliability of each component of the composite. Generalizability theory, classical test theory, and IRT all have associated procedures for estimating reliability of composites.

Subpopulations

Generalizability coefficients, reliability coefficients, decision consistency coefficients, and average standard errors of measurement require specification of the population. These indices can differ for different populations and subpopulations. In situations where there is a number of important subpopulations, it is crucial to estimate these coefficients for each subpopulation.

Data Collection

The estimation of reliability indices and standard errors of measurement requires that data be collected. Ideally, these data include variations of the characteristics that change from one replication of the measurement procedure to another. For example, data collection would include each examinee taking multiple sets of tasks, and multiple raters would score these constructed-response tasks. However, such data collection might not be feasible in an operational testing program. In this case, alternate data collection designs treated outside of the operational context might be used to estimate reliability indices.

Using Reliability Information

After reliability information is estimated, the information can be used to help design the measurement procedure. Such information can also be used to document the amount of error in test scores and to report information about error to test users.

One of the major uses of reliability information is to decide on the number of tasks that are necessary to achieve a particular level of reliability for a given population of examinees. For example, suppose that the reliability coefficient for a measurement procedure is desired to be

.90 or greater. Using reliability information, the number of tasks that need to be included in an assessment to achieve the stated level can be estimated.

Another use of reliability information is to decide on the number of raters to be used to a particular level of reliability for a given population of examinees and group of raters. Generalizability theory allows joint consideration of the numbers of tasks and numbers of raters to use to achieve a particular level of generalizability or reliability.

Reliability information can also be used to calculate average and conditional standard errors of measurement, which can then be used to help test users interpret test scores. Such information also can be used to document, in technical documentation, the amount of different sources of error on assessments.

Components and Scores for PARCC and SBAC Assessments

In this section, the components and scores of the PARCC and SBAC proposed assessment systems that are most important to considerations of reliability are summarized. The PARCC assessments are described first, followed by the SBAC assessments.

PARCC

The following description is based on material taken from the PARCC (2010) application, which provides plans for the assessments, and from ETS (2010).

PARCC assessments. Each PARCC English and language arts (ELA) and mathematics (Math) assessment is administered in Grades 3–11. Brief descriptions of the assessments follow.

- ELA-1 is administered following ≈25% of instruction and contains 1–2 extended constructed-response tasks.
- ELA-2 is administered following ≈50% of instruction and contains 1–2 extended constructed-response tasks.

- ELA-3 is administered following ≈75% of instruction and contains 1 multiday performance task.
- ELA-EOY is administered following ≈90% of instruction and contains 45–65 tasks.
- ELA-4 is administered following ELA-3 and contains 1 speaking and listening task.
- ELA Summative Weighted Composite Score is calculated from the ELA-1, ELA-2, ELA-3, and ELA-EOY components.
- Math-1 is administered following ≈25% of instruction and covers 1–2 topics with 2 brief constructed-response tasks and 1 extended constructed-response task per topic.
- Math-2 is administered following ≈50% of instruction and covers 1–2 topics with 2 brief constructed-response tasks and 1 extended constructed-response task per topic.
- Math-3 is administered following ≈75% of instruction and contains 1 extended performance task.
- Math-EOY is administered following ≈90% of instruction and contains 40–50 tasks.
- Math Summative Weighted Composite Score is calculated over Math-1, Math-2, Math-3, and Math-EOY components.

PARCC scores for each student. For through-course assessments components ELA-1, ELA-2, ELA-3, ELA-EOY, Math-1, Math-2, Math-3, and Math-EOY, an achievement (raw) score and a readiness performance level are reported for each student. For ELA-4, only an achievement (raw) score is reported. Following completion of all assessment components, the ELA and Math Summative Weighted Composite Scores are calculated and reported as scale scores. Subscale scores, performance category classifications, and growth indices also are provided.

For through-course assessments components ELA-1, ELA-2, ELA-3, Math-1, Math-2, and Math-3, scores are used to monitor student progress, inform instructional decisions, signal

interventions for individuals and are aggregated to assess teacher and curriculum effectiveness. ELA-4 scores are used only for instructional purposes. For through-course assessment component ELA-EOY and Math-EOY, scores are used to determine whether students are on track and have met the standards for the year.

PARCC scores for aggregates (e.g., classroom, school, district). All scores calculated at the individual level are aggregated, except for ELA-4. The aggregated scores are used to measure district, school, and teacher effectiveness, student proficiency, and student growth.

SBAC

The following description is based on material taken from the SBAC (2010) application, which provides plans for the assessments, and from ETS (2010).

SBAC assessments. Brief descriptions of the SBAC assessments follow.

- ELA Reading Interim/Benchmark Assessment, with multiple testing opportunities per year, contains selected-response and constructed-response tasks.
- ELA Writing, Listening, and Speaking and Language Interim/Benchmark Assessment, with multiple testing opportunities per year, contains selected-response and constructed-response tasks.
- Mathematics Interim/Benchmark Assessment, with multiple testing opportunities per year, contains selected-response and constructed-response tasks.
- ELA Reading Summative Assessment contains selected-response tasks, constructed-response tasks, and 1 performance event task. Students have the option to retake the test once with a different set of tasks. The performance events and other tasks are administered within a 12-week window near the end of the school year.
- ELA Summative Writing, Listening, and Speaking and Language contains selected-response tasks, constructed-response tasks, and 1 performance event task. Students have the option to retake the test once with a different set of tasks. The

performance events and other tasks are administered within a 12-week window near the end of the school year.

- Summative Mathematics contains selected-response tasks, constructed-response tasks, and 2 performance event tasks. Students have the option to retake the test once with a different set of tasks. The performance events and other tasks are administered within a 12-week window near the end of the school year.

SBAC scores for each student. For each of the interim/benchmark assessments, a scale score is reported to indicate achievement and growth, and content cluster scores are reported to indicate growth and inform instruction. For the summative assessments, a scale score on a vertical scale is reported to indicate achievement and growth.

SBAC scores for aggregates (e.g., classroom, school, district). For both interim/benchmark and summative assessments, scores are aggregated at the class, school, district, and state levels.

SBAC assessment option. SBAC "will conduct studies to determine whether distributed summative assessments (a series of tests taken across the school year) are sufficiently valid, reliable, and comparable to the … EOY assessments to be offered as an alternative to the current EOY assessment" (ETS, 2010, p. 13).

Unique Challenges in Assessing Reliability of Scores for PARCC and SBAC Assessments and Potential Solutions

Challenges for assessing reliability and for achieving adequate reliability with the PARCC and SBAC assessments are considered in this section. These challenges involve the following:

- Reliability of assessment components containing all or mainly constructed-response tasks

- Reliability of scores used for accountability purposes that are composites of other scores
- Decision consistency for performance levels
- Reliability of all of the different types of scores that will be used with these assessments including subscores, cluster scores, scores for aggregates (e.g., classrooms, schools, districts, states), and growth indices
- Reliability for important subpopulations

In the discussion of each of these challenges, potential solutions are also described.

Scores on Constructed-Response Components

Both PARCC and SBAC make substantial use of constructed-response tasks. Both consortia plan to do at least some of the scoring of the constructed-response tasks using computer-based automated scoring. The extensive use of constructed-response tasks creates challenges, including how to estimate reliability of scores over a small number of constructed-response tasks, how to ensure that the scores are of adequate reliability, and how to estimate reliability when responses are scored using computer-based automated scoring.

Estimating reliability. One of the benefits of using selected response tasks is that a test form that contains many tasks can be administered to examinees in a single administration. Subsets of the tasks included in the test form often can be considered to adequately represent the construct of interest. In this case, these subsets can be viewed as replications of the task characteristic of the measurement procedure, and this information can be used to estimate reliability of scores on the assessment for a test that contains multiple subsets of representative tasks.

Constructed-response tasks typically are more time-consuming for examinees than are selected-response tasks. Often, only a small number of such tasks can be administered within reasonable time limits. When tests consist of few constructed-response tasks, it might be difficult to consider multiple subsets of tasks within an assessment to represent the construct

of interest. In the extreme case where there is only one constructed-response task that has a single score, it is impossible to use information on multiple subsets of representative tasks to estimate reliability.

Many of the components of the PARCC assessments contain a small number (1 or, at most, 2) extended constructed-response or performance tasks. These include ELA-1, ELA-2, ELA-3, Math-1, Math-2, and Math-3. Scores on these components are to be used to monitor student progress, inform instructional decisions, signal interventions for individuals and are aggregated to assess teacher and curriculum effectiveness and accountability. The SBAC assessments also contain small numbers of constructed-response tasks, and the summative assessments contain 1 performance task, although it appears that scores might not be calculated for the constructed-response portions of the test. Because the constructed-response components have so few tasks, it may be difficult to adequately estimate reliability using data from operational administrations.

One way to address the issue of estimation of reliability is to conduct pilot studies during the development of the PARCC and SBAC assessments. One design that could be used involves administering at least two forms of the constructed-response assessments to examinees. Two or more raters would score the examinee responses to each form. In generalizability theory terminology, such a design would be a *persons (p) x tasks (t) x raters (r)* or *p x t x r* design (Haertel, 2006, p. 90). Such a design could be used to estimate reliability for assessments that contain different numbers of tasks scored by different numbers of raters. When analyzed using generalizability theory, this design can be used to provide a comprehensive analysis of errors of measurement including estimation of components of error variation that involve interactions of persons with tasks and raters. The results from such a study could be used (a) to assess whether the scores are sufficiently reliable for the desired uses and (b) to plan for modifications to the assessment procedures that include using different numbers or types of tasks, changing the scoring rubrics, and changing the training of raters.

Increasing reliability. PARCC and SBAC plan to use small numbers of constructed-response tasks. Using few tasks might lead to inadequate reliability for making educationally important decisions based on scores on the PARCC and SBAC assessments.

Reliability typically increases as tests become longer. Thus, scores on tests that contain many selected-response tasks often are quite reliable. One strategy for increasing the reliability of the components for the constructed-response tasks is to develop constructed-response tasks that consist of a number of separately scored components. With many separately scored components, it might be possible to treat the constructed-response task as a series of tasks with scores that could be treated independently and where subsets of components on a particular form could be treated as replications of the measurement procedure. Because there are more independent components that contribute to the score, this sort of test structure could lead to a total score that is more reliable than when a single, holistic score is used with one component of the assessment.

Automated scoring. Both the PARCC and SBAC assessment systems intend to make extensive use of computer-based automated scoring. In such systems, raters typically score a sample of responses that are used to calibrate the automated system. An independent set of responses is then scored by two raters and by the automated system. According to Drasgow, Luecht, and Bennett (2006), “if the automated scores agree with the human judges to about the same degree as the human judges agree among themselves, the automated system is considered to be interchangeable with the scores of the typical judge” (p. 497). Thus, the automated scoring systems often are intended to mirror human raters. Given this intent, it might be reasonable to estimate reliability of scores that include automated scoring by estimating reliability of scores based on human judges.

However, it would be preferable to use procedures for directly estimating the reliability of scores that include the use of automated scoring, although such procedures have yet to be developed. One possible approach would be to assess individuals with replications of different

tasks scored using automated scoring and to assess the reliability of the scores over the tasks. The development of procedures for assessing the reliability of scores that are based on automated scoring systems is an important area for further research.

Weighted Composites

Weighted composite scores are used in both the PARCC and SBAC assessments. These composites are calculated over assessments given at different times and over assessments containing different task types. According to Standard 2.7 in *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), "When subsets of items within a test are dictated by the test specifications and can be presumed to measure partially independent traits or abilities, reliability estimation procedures should recognize the multifactor character of the instrument" (p. 33).

Scores from assessments administered at different times. For the PARCC assessments, the ELA and Math Summative Weighted Composite scores are calculated over assessments that are given at four different times during the school year. Because students are tested at different times, it is not reasonable to assume that students' proficiency is constant at the various times that the assessments are given. For example, it would be unreasonable to fit a single unidimensional IRT model to the examinee item responses across all components of the assessment. Instead, a psychometric approach can be taken that estimates error variability for each component separately. To estimate reliability of scores for the Summative Weighted Composites, reliability of scores for each component of the composite can be estimated. By assuming that measurement errors are independent across the components, an estimate of error variance can be found by taking a weighted sum of the error variances for each component. A reliability coefficient can be estimated as 1 minus the ratio of error variance to total composite score variance (Haertel, 2006, p. 76).

Scores based on mixed formats. With both the PARCC and SBAC assessments, composite scores are calculated over scores from different task types. Evidence exists that the

skills assessed by the different task types in these kinds of mixed-format tests are often distinct (Rodriguez, 2003). Although it might be possible in some cases to use a unidimensional model to assess reliability of mixed-format tests (Wainer & Thissen, 1993), the use of such a model can lead to different, and possibly inaccurate, estimation of reliability (Kolen & Lee, in press). In general, it is preferable to assume that the different task types assess different proficiencies, to assess reliability of each component separately, and to use the procedure for estimation of reliability of composite scores discussed in the previous paragraph.

The SBAC assessments contain portions of tests that are administered adaptively. Estimation of reliability for the SBAC assessments could proceed by estimating reliability for each component separately and then combining the error variances for the components in much the same way as was suggested in the previous two paragraphs.

Weights. The weights for each component of composites like those proposed for the PARCC and SBAC assessments can have a substantial effect on the reliability of the composite (Kolen & Lee, in press; Wainer & Thissen, 1993). Because they typically are based on few tasks, scores on constructed-response task components often have relatively greater error variability than scores on selected-response task components per unit of testing time. However, for practical reasons, it may be important to make sure that the constructed-response components are weighted highly. If components with relatively larger error variability receive large weights, then it is possible for the composite scores to have lower reliability than some of the components. In developing weights, it is often necessary to balance the practical need to give substantial weight to the constructed-response task components and the psychometric need to have composite scores that are adequately reliable. Fortunately, often a range of weights can be used that meet both criteria well (Kolen & Lee; Wainer & Thissen). When developing weights for the PARCC ELA and Math Summative Weighted Composite Scores and the SBAC Summative scores, both the practical issue of providing sufficient weight to the constructed-response tasks and the need to have composite scores that are of adequate reliability should be considered.

Achievement Level Classifications

PARCC and SBAC indicate that achievement levels will be associated with the composite scores. It will be important to assess decision consistency with these achievement levels and to ensure that the decision consistency is of an appropriate magnitude for the decisions to be made. This recommendation is reinforced by AERA, APA, and NCME (1999) Standard 2.15 which states:

> … when a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument. (p. 35)

Other Scores

PARCC refers to the use of subscores, but without much detail. SBAC refers to the use of content cluster scores to be reported with the interim/benchmark and assessments to indicate growth and inform instruction. It will be important to the reliability of the subscores and to only report subscores that are sufficiently reliable to support the interpretations that are made.

Both PARCC and SBAC refer to aggregation of scores to levels such as the classroom, school, district, and state. Models for assessing reliability at these different levels of aggregation (see Brennan, 2001; Haertel, 2006) should be applied to the aggregated scores to estimate reliability.

Both PARCC and SBAC refer to using the scores from these assessments in growth models. Such models would include growth across grades as well as possibly within grade. The growth indices likely will be estimated at the aggregate level (classroom, school, district, state). The reliability of growth indices at the different levels of aggregation should be estimated. Estimation of reliability of growth indices will necessarily incorporate the measurement error of scores at the different points in time that are used in the calculation of the growth indices.

SBAC refers to the use of a vertical scale, but without much detail about the vertical scale. Reliability information for scores on the vertical scale should be provided. In addition, if growth for individuals or aggregates is to be assessed using the differences between scores on the vertical scale that reflect performance at different grades or times, then reliability information of such difference scores will need to be estimated and provided.

All Students

Many of the tasks that are included in the PARCC (2010) and SBAC (2010) applications involve the use of complex stimuli and require open-ended written responses from examinees. The use of such materials include "challenging performance tasks and innovative, computer-enhanced items that elicit complex demonstrations of learning . . ." (PARCC, 2010, p. 7) and "that reflect the challenging CCSS [Common Core State Standards] content, emphasizing not just students' 'knowing,' but also 'doing'" (SBAC, 2010, p. 37). An important open question is whether such tasks can adequately assess students performing at all levels of the achievement continuum without or with accommodations. It seems possible that such assessments could pose particular problems for students who are English language learners and for students with certain types of disabilities.

According to Standard 2.11 (AERA, APA, & NCME, 1999):

> … if there are generally accepted theoretical or empirical reasons for expecting that reliability coefficients, standard errors of measurement, or test information functions will differ substantially for various subpopulations, publishers should provide reliability data as soon as feasible for each major population for which the test is recommended. (p. 34)

Because of the complexity of the tasks used with the proposed PARCC and SBAC assessments, it seems likely that there will be reliability differences across subgroups. For this reason, reliability of scores for different subgroups should be estimated, including gender,

racial/ethnic, disability, and English language learner subgroups. It is important that all of the scores reported (e.g., scores of different through-course components, composite scores, subscores, cluster scores, aggregated scores, and growth indices) are reliable for all subgroups of students.

The assessment of reliability of all scores for all important subgroups of students should be accomplished during the development of the assessments. Inadequate reliability for any subgroup should lead to possible modification of the assessments. Reliability of all scores for all important subgroups also should be accomplished following the development of the assessments in order to document that for all subgroups, all of the scores have adequate reliability for their intended purposes.

Challenges and Recommendations

The proposed PARCC and SBAC are quite complex, using a variety of item types, having components given at different times during the year, reporting a large number of different types of scores, and being appropriate for a wide range of students. Due to the complexity of the assessments, it will be necessary to conduct a variety of pilot studies during the development of the test to assess reliability of the scores on the assessments so that the scores on the operational assessments will be sufficiently reliable for their intended purposes. The challenges identified and the recommendations made in this paper are summarized in Table 1 and in the next section.

Table 1. Recommendations and Challenges for Through-course Assessments

Challenge 1. The use of a small number of constructed-response tasks in various components of the proposed PARCC and SBAC assessments makes it difficult, if not impossible, to adequately estimate reliability of the components using data from operational administrations.	**Recommendation 1.** Conduct pilot studies during the development of the PARCC and SBAC assessments using a *p x t x r* design that allow for the estimation of reliability of the assessments using different numbers of tasks and raters. The results from these pilot studies can be used to refine the assessment tasks, rating procedures, and assessment design so that scores on the constructed-response components of the assessments are of adequate reliability.
Challenge 2. The use of a small number of constructed-response tasks in various components of the proposed PARCC and SBAC assessments might lead to inadequate reliability of scores on the constructed-response components of these assessments.	**Recommendation 2.** To the extent possible, develop constructed-response tasks that consist of a number of separately scored components that could lead to more reliable scores than if one holistic score is used with each constructed-response task.
Challenge 3. Methods for assessing reliability of scores on tests where constructed responses are scored using automated scoring have not been fully developed.	**Recommendation 3.** (a) Estimate reliability of scores based on human judgment and use these to represent reliability of scores based on automated scoring. (b) Conduct research on procedures for assessing the reliability of scores that are based on automated scoring systems.
Challenge 4. For both proposed PARCC and SBAC assessments, components of the assessments are administered at different times, so it is not reasonable to assume that students' proficiency is constant over the various times.	**Recommendation 4.** Use psychometric methods that do *not* assume that student proficiency is constant over the various times of administration. Instead, estimate reliability for each component separately and use psychometric procedures that are designed to assess reliability for composite scores.

Challenge 5. For both proposed PARCC and SBAC assessments, composite scores are calculated over components of the assessments that consist of different task types, such as constructed-response and selected-response tasks.	**Recommendation 5.** Use psychometric methods that do *not* assume that student proficiency is the same over the different task types. Instead, estimate reliability for each component separately and use psychometric procedures that are designed to assess reliability for composite scores.
Challenge 6. The weights used for each component of composites can have a substantial effect on the reliability of the composite.	**Recommendation 6.** In developing the weights for each component of weighted composites, balance the practical need to give substantial weight to the constructed-response task components and the psychometric need to have weighted composite scores that are adequately reliable for their intended purposes.
Challenge 7. A variety of scores will be reported for the assessments, including scores on components, composite scores, achievement levels, subscores, content cluster scores, scores for aggregates, growth indices for individuals and aggregates, and vertical scale scores.	**Recommendation 7.** (a) During development of the assessments, conduct pilot studies to estimate the reliability of the scores and modify the assessments, where needed, to achieve adequate score reliability. (b) Assess the reliability of each of the different types of scores for the assessments that are administered operationally.
Challenge 8. The use of complex stimuli that require open-ended written responses could lead to assessment tasks that do not provide reliable scores for various examinee groups, including English language learners and students with various disabilities.	**Recommendation 8.** (a) During development of the assessments, conduct pilot studies to estimate reliability for each subgroup (including English language learners and students with various disabilities) and modify the assessments, where needed, to achieve adequate reliability for all students. (b) Assess reliability for each subgroup for the assessments that are administered operationally.

Challenge and Recommendation 1: Assessing Reliability for Scores on Constructed-Response Tasks

Challenge 1. The use of a small number of constructed-response tasks in various components of the proposed PARCC and SBAC assessments makes it difficult, if not impossible, to adequately estimate reliability of the components using data from operational administrations.

Recommendation 1. Conduct pilot studies during the development of the PARCC and SBAC assessments using a *p x t x r* design that allow for the estimation of reliability of the assessments using different numbers of tasks and raters. The results from these pilot studies can be used to refine the assessment tasks, rating procedures, and assessment design so that scores on the constructed-response components of the assessments are of adequate reliability.

Challenge and Recommendation 2: Increasing Reliability for Scores on Constructed-Response Components

Challenge 2. The use of a small number of constructed-response tasks in various components of the proposed PARCC and SBAC assessments might lead to inadequate reliability of scores on the constructed-response components of these assessments.

Recommendation 2. To the extent possible, develop constructed-response tasks that consist of a number of separately scored components that could lead to more reliable scores than if one holistic score is used with each constructed-response task.

Challenge and Recommendation 3: Reliability for Scores on Constructed-Response Components Scored Using Automated Scoring

Challenge 3. Methods for assessing reliability of scores on tests where constructed responses are scored using automated scoring have not been fully developed.

Recommendation 3. (a) Estimate reliability of scores based on human judgment and use these to represent reliability of scores based on automated scoring. (b) Conduct research on procedures for assessing the reliability of scores that are based on automated scoring systems.

Challenge and Recommendation 4: Reliability for Scores on Assessments Consisting of Components Administered at Different Times

Challenge 4. For both proposed PARCC and SBAC assessments, components of the assessments are administered at different times, so it is not reasonable to assume that students' proficiency is constant over the various times.

Recommendation 4. Use psychometric methods that do *not* assume that student proficiency is constant over the various times of administration. Instead, estimate reliability for each component separately and use psychometric procedures that are designed to assess reliability for composite scores.

Challenge and Recommendation 5: Reliability for Scores on Mixed-Format Assessments

Challenge 5. For both proposed PARCC and SBAC assessments, composite scores are calculated over components of the assessments that consist of different task types, such as constructed-response and selected-response tasks.

Recommendation 5. Use psychometric methods that do *not* assume that student proficiency is the same over the different task types. Instead, estimate reliability for each component separately and use psychometric procedures that are designed to assess reliability for composite scores.

Challenge and Recommendation 6: Weighting Scores Across Components

Challenge 6. The weights used for each component of composites can have a substantial effect on the reliability of the composite.

Recommendation 6. In developing the weights for each component of weighted composites, balance the practical need to give substantial weight to the constructed-response task components and the psychometric need to have weighted composite scores that are adequately reliable for their intended purposes.

Challenge and Recommendation 7: Assessing Reliability of All Scores

Challenge 7. A variety of scores will be reported for the assessments, including scores on components, composite scores, achievement levels, subscores, content cluster scores, scores for aggregates, growth indices for individuals and aggregates, and vertical scale scores.

Recommendation 7. (a) During development of the assessments, conduct pilot studies to estimate the reliability of the scores and modify the assessments, where needed, to achieve adequate score reliability. (b) Assess the reliability of each of the different types of scores for the assessments that are administered operationally.

Challenge and Recommendation 8: Assessing Reliability for All Students

Challenge 8. The use of complex stimuli that require open-ended written responses could lead to assessment tasks that do not provide reliable scores for various examinee groups, including English language learners and students with various types of disabilities.

Recommendation 8. (a) During development of the assessments, conduct pilot studies to estimate the reliability for each subgroup (including English language learners and students with various disabilities) and modify the assessments, where needed, to achieve adequate reliability for all students. (b) Assess score reliability for each subgroup for the assessments that are administered operationally.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*(4), 364–375.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: Praeger.

Council of Chief State School Officers, & National Governors Association Center. (2010). *Common core state standards initiative*. Retrieved from http://www.corestandards.org

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: Praeger.

Educational Testing Service. (2010). *Coming together to raise achievement. New assessments for the common core state standards.* Retrieved from http://k-12center.com/rsc/pdf/Assessments_for_the_Common_Core_Standards.pdf

Haertel, E. H. (2006).Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Kolen, M. J., & Lee, W. C. (in press). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practices*.

Partnership for Assessment of Readiness for College and Careers. (2010). *Application for the Race to the Top Comprehensive Assessment Systems Competition*. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163–184.

SMARTER Balanced Assessment Consortium. (2010). *Race to the Top assessment program application for new grants*. Retrieved from http://www.k12.wa.us/SMARTER/RTTTApplication.aspx

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103–118.

Yen, W., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger.