Invitational Research Symposium on
Through-Course
Summative Assessments

SCALING AND LINKING THROUGH-COURSE SUMMATIVE ASSESSMENTS

Rebecca Zwick and Robert J. Mislevy

Educational Testing Service

March 2011

Center for K – 12 Assessment
& Performance Management at ETS

Scaling and Linking Through-Course Summative Assessments

Rebecca Zwick and Robert J. Mislevy
Educational Testing Service

Executive Summary

A new entry in the testing lexicon is the *through-course summative assessment*, defined in the notice inviting applications for the Race to the Top program as "an assessment system component or set of assessment system components that is administered periodically during the academic year. A student's results from through-course summative assessments must be combined to produce the student's total summative assessment score for that academic year" (Department of Education, 2010, p. 18178). A key feature of through-course summative assessments (TCSAs), then, is that, unlike interim or benchmark assessments, TCSAs are specifically intended to be combined into a summary score that is to be used for accountability purposes.

A reading of the notice inviting applications and the consortium proposals reveals a number of properties the TCSAs are intended to have: They are expected to include multidimensional content and complex performance tasks, and they must be available in multiple equivalent test forms. They must accommodate students who vary in their patterns of curricular exposure and must allow for the meaningful measurement of growth over the course of an academic year.

The TCSAs must also serve as the basis of both individual and group proficiency estimates. Because results will include percentages of students at each of several achievement levels, proficiency distributions (not merely averages) must be correctly estimated for each grade level and for all relevant student groups.

Because the inferential demands on the TCSAs are extensive, a complex analysis model will likely be required. We propose a model that is similar to those used by the National

Assessment of Educational Progress (NAEP), the Programme in International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS). The model includes an item response theory (IRT) component, which characterizes the properties of test items, and a population component, which summarizes the characteristics of proficiency distributions. An advantage of the model is that a body of literature already exists regarding its statistical properties. A disadvantage is its conceptual and computational complexity.

We note, however, that the use of a complex analysis model does not imply that the reporting model be complex as well. In fact, we suggest that results be expressed in terms of expected performance on a particular set of tasks—a market-basket reporting strategy. This kind of reporting metric tends to be easily interpretable and also lends itself well to the measurement of growth.

We describe conditions under which simplifications in analysis might be possible, and we recommend that these potential simplifications be investigated in the pilot and field test phases of the Race to the Top venture. One strategy worthy of exploration is to use relatively unconstrained test forms for certain purposes, such as informing classroom instruction, while using more constrained test forms, administered to only a sample of students, for other inferences, such as comparisons of schools, districts, and states. These more rigorously constructed forms, which would consist of parallel tests with controlled mixtures of machine-scoreable item types, could be seeded into test administrations. Basing the primary reporting scale on only the more constrained forms would allow the use of more flexibly created assessments, including extended performance tasks, for instructional purposes. A dual strategy of this kind could also facilitate the use of a simpler IRT model for the primary reporting scale.

In general, we stress the importance of recognizing the tradeoffs between inferential demands and procedural simplicity. The more demands that are made of the scaling model for the TCSAs, the more complex the model needs to be. As demands are reduced, simpler approaches become more feasible.

Scaling and Linking Through-Course Summative Assessments

Rebecca Zwick and Robert J. Mislevy[1]
Educational Testing Service

A new entry in the testing lexicon is the *through-course summative assessment*, defined in the notice inviting applications for the Race to the Top Race to the Top program as "an assessment system component or set of assessment system components that is administered periodically during the academic year. A student's results from through-course summative assessments must be combined to produce the student's total summative assessment score for that academic year" (Department of Education, 2010, p. 18178). A key feature of through-course summative assessments (TCSAs), then, is that, unlike interim or benchmark assessments, TCSAs are specifically intended to be combined into a summary score that is to be used for accountability purposes.

Two consortia of states have received Comprehensive Assessment Systems grants under Race to the Top: The SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Although the SBAC application (SBAC, 2010) stated only that the implementation of TCSAs will be studied, the PARCC application (PARCC, 2010) included detailed proposals about the use of TCSAs.

This paper addresses the scaling and linking issues associated with TCSAs. In the next section of the paper, Inferences Intended From Through-Course Summative Assessments and Their Implications for Assessment Design, we outline the kinds of inferences that are intended to be made using TCSAs and the resulting implications for assessment design. In the section titled A General Model for Scaling, Linking, and Reporting Through-Course Summative

[1] We appreciate the comments of Brent Bridgeman, Daniel Eignor, Shelby Haberman, Steven Lazer, Andreas Oranje, and Matthias von Davier. All positions expressed in this paper are those of the authors and not necessarily those of Educational Testing Service.

Assessments, we outline a general psychometric framework for scaling, linking, and reporting TCSA results. Some challenges in simultaneously supporting all of the inferences listed in the previous section are noted. In the Illustration of Individual Student Proficiency Estimation section, we show how the model could be applied to obtain individual test scores in a specific assessment situation and how results from through-course assessment occasions might be combined. The General Procedures for Estimating Population Characteristics section describes procedures for the estimation of population characteristics and provides more detail about computational issues. Finally, in last section, Discussion and Recommendations, we summarize our presentation, discuss circumstances under which the model could be simplified, and offer some recommendations.

Inferences Intended From Through-Course Summative Assessments and Their Implications for Assessment Design

According to the notice inviting applications (Department of Education, 2010), assessments developed under the Comprehensive Assessment Systems grants must "measure student knowledge and skills … in mathematics and English language arts in a way that … provides an accurate measure of student achievement across the full performance continuum and an accurate measure of student growth over a full academic year or course"(p. 18171). As mentioned earlier, it is also required that "a student's results from through-course summative assessments be combined to produce the student's total summative assessment score for that academic year" (Department of Education, 2010, p. 18178).

The SBAC application makes only a brief reference to through-course assessment, stating that the consortium will investigate "the reliability and validity of offering States an optional distributed summative assessment as an alternative for States to the administration of the summative assessment within the fixed 12-week testing window…" The proposal goes on to say that "[t]he scores of these distributed assessments would be rolled up (along with students'

scores on the performance events) to make the overall decision about students' achievement…" (SBAC, 2010, p. 43).[2]

The PARCC application includes a substantial amount of information about its proposed use of TCSAs: The partnership proposes to "distribute through-course assessments throughout the school year so that assessment of learning can take place closer in time to when key skills and concepts are taught and states can provide teachers with actionable information more frequently" (PARCC, 2010, p. 7). The PARCC application further notes the following:

> The Partnership's summative assessment system will consist of four components distributed throughout the year in [English language arts (ELA)]/literacy, three components distributed throughout the year in mathematics, and one end-of-year component in each subject area. The first three through-course components in ELA/literacy and mathematics will be administered after roughly 25 percent, 50 percent and 75 percent of instruction. The fourth through-course component in ELA/literacy, a speaking and listening assessment, will be administered after students complete the third component … [but will not contribute to the combined summative score]. The end-of-year components in each subject area will be administered after roughly 90 percent of instruction. The assessment system will include a mix of constructed-response items; performance tasks; and computer-enhanced, computer-scored items. (PARCC, 2010, pp. 44–45)

In particular, in both math and English language arts, students will be expected to "participate in an extended and engaging performance-based task" during their third TCSA of the school year" (PARCC, 2010, p. 36).

The PARCC application also describes the consortium's reporting plans:

[2] According to supplementary information obtained from SBAC (N. Doorey, personal communication, December 12, 2010), the TCSAs would consist of "a series of computer-adaptive assessments…given across the school year." The assessments would include "perhaps 20 to 40 items of multiple item types that can be electronically scored" and possibly additional items requiring other scoring procedures.

> Information will be provided at the individual student level and, as appropriate, aggregated and reported at the classroom, school, district, state, and Partnership level. Subgroup data, to be reported at the school level and higher, will include ethnic group, economically disadvantaged, gender, students with disabilities and English learners (p. 63).

Also, reported results will not be restricted to means. PARCC will establish four performance level classifications and will report the percentage of students at each performance level (PARC, 2010, p. 63).

Based on the requirements of the notice inviting applications (Department of Education, 2010) and the PARCC proposal's intended usage, we can infer the following list of properties of TCSAs, which must be accommodated by the data analysis model:

1. Each TCSA is associated with a segment of the curriculum. These curricular segments collectively represent multiple content and skill areas and hence, any summary scale formed from the TCSAs is potentially multidimensional.
2. The requirement to measure *growth* over the course of an academic year implies that the assessment must be capable of measuring some underlying construct (such as "knowledge of grade 3 math") or at least some meaningful aggregate of a set of components (e.g., 20% numerical operations, 40% algebra, and 40% geometry). Without this overarching conceptualization of a subject area, we would be required to measure growth by comparing, say, performance in numerical operations at Time 1 to performance in algebra at Time 2, which is clearly not a meaningful enterprise.
3. The requirement to measure academic-year growth, interpreted strictly, also implies that a measurement is needed at the beginning of the school year. (The PARCC proposal does not explicitly include such a measurement.)

4. Since schools are not constrained to any particular curricular order, the TCSAs themselves could be given in different orders across schools, districts, and states. Furthermore, even schools that use the same curricular order may take differing amounts of time to cover a particular area, implying that the time of testing could vary.
5. Because of the risk of security breaches (which is especially high due to varying curricular orders and testing times), an assessment based on a fixed set of items is clearly not feasible, even within a school year. Multiple forms of each TCSA must be available.
6. Because complex performance tasks are to be included, the assessment can be expected to contain some items that are scored on an ordinal scale (not simply right-wrong), and possibly some clusters of dependent items.
7. Because subgroup results are to be reported, issues relevant to the unbiased estimation of subpopulation characteristics must be considered. Because results include percentages of students at each of several achievement levels, proficiency distributions (not merely means) must be correctly estimated for the population and all relevant subpopulations.
8. Because the TCSAs are meant to provide actionable information, it can be inferred that tasks are intended to be sensitive to instruction. That is, the difficulty of an item can be expected to change notably when students are exposed to the instructional activities that target the relevant proficiencies. This implies that difficulties of tasks can be expected to vary relative to one another across curricular segments, and possibly within curricular segments as well.

A General Model for Scaling, Linking, and Reporting Through-Course Summative Assessments

The objective of this section is to lay out an inclusive psychometric framework for supporting the inferences listed in the notice inviting applications (Department of Education,

2010) and in the consortium proposals. The proposed framework takes into account Properties 1 through 8 of TCSAs listed in the previous section and builds on research and experience from previous large-scale assessments. A latent variable model is presented in order to allow for variations in test design and to facilitate linking results across forms, moving items into and out of the pool and integrating results from different forms into a common reporting scheme.

To represent the testing situation outlined in the previous section, we define, for student *i,* a vector of item responses, $\mathbf{x}_i$, a vector of curricular variables, $\mathbf{c}_i$, and a vector of demographic variables, $\mathbf{d}_i$. The number of elements in $\mathbf{x}_i$ is the number of items on all TCSAs for a given academic year, combined. The curricular variables $\mathbf{c}_i$ include information about the nature and order of the curriculum to which student *i* was exposed, as well as information about when he or she was tested. The vector $\mathbf{d}_i$ contains demographic information, such as student *i*'s ethnicity and gender and, as explained below, must include all categorizations for which subpopulation distributions are to be estimated. We focus on assessing performance at a single testing occasion and then briefly note additional considerations in measuring growth and combining TCSA results across occasions.

We wish to make inferences about Θ, an unobserved proficiency variable that is expected to be multidimensional. In the Illustration of Individual Student Proficiency Estimation section, for example, Grade 3 math is assumed to comprise numerical operations, algebra, and geometry. (Empirical analyses may indicate that less transparent models are needed in some applications.) We want to estimate population characteristics for various demographic groups, such as girls, African-American students, or English language learners, and also want to estimate the proficiencies of individual students.

Using the Bayesian framework of Mislevy (1985), the basic model can be expressed as

$$p(\Theta|\mathbf{x}_i,\mathbf{c}_i,\mathbf{d}_i) \propto P(\mathbf{x}_i|\Theta,\mathbf{c}_i,\mathbf{d}_i)\ p(\Theta|\mathbf{c}_i,\mathbf{d}_i) = P(\mathbf{x}_i|\Theta)\ p(\Theta|\mathbf{c}_i,\mathbf{d}_i) \quad (1)$$

The term to the left of the *is proportional to* ($\propto$) sign, $p(\Theta|\mathbf{x}_i,\mathbf{c}_i,\mathbf{d}_i)$, represents the distribution of Θ given the observed item responses and background data, referred to as the

posterior distribution of Θ for student *i*. $P(\mathbf{x}_i \mid \Theta, \mathbf{c}_i, \mathbf{d}_i)$ represents the distribution of the item responses, given the values of Θ, $\mathbf{c}_i$, and $\mathbf{d}_i$, that is, the (presumably multidimensional) item response model. Multiple-category models, rater-effects models, and other more complex models are subsumed in the notation. Once the data are collected, this term represents the likelihood function for Θ.

It is an important feature of our model that effects of instruction are assumed to be captured in the term $p(\Theta \mid \mathbf{x}_i, \mathbf{c}_i, \mathbf{d}_i)$ and in the multidimensional IRT (MIRT) model, as described more fully below. We wish to invoke the typical IRT assumption that, conditional on Θ item response functions do not depend on background variables and do not vary over testing occasions, allowing us to simplify $P(\mathbf{x}_i \mid \Theta, \mathbf{c}_i, \mathbf{d}_i)$ to $P(\mathbf{x}_i \mid \Theta)$. As discussed further in the Modeling Growth Across Occasions subsection, we do not propose assuming conditional independence by fiat, but instead advocate using observed patterns in the data to inform construction of the IRT model. The goal is to structure the model so that conditional independence is approximated well enough to support the desired inferences.[3] As discussed in the Discussion and Recommendations section, it may be possible to mitigate the complexity

[3] As a somewhat simplistic example of how a multidimensional IRT model (with invariant item parameters) could capture the effects of instructional sensitivity, suppose that we have a two-dimensional model with θ_1 = proficiency in aspects of the relevant content that are less instructionally sensitive and θ_2 = proficiency in aspects that are more instructionally sensitive. Under a typical 2-dimenstional MIRT model, the probability of correct response will depend on the term $g_1\theta_1 + g_2\theta_2 + k$, where g_1 and g_2 are item discriminations (which can also be viewed as factor loadings) and *k* is the usual item difficulty parameter. (A guessing parameter could be included as well, but is not relevant here.) In this model, instructionally sensitive items have high g_2 values. After instruction, a student's θ_1 may undergo little change, but his θ_2 increases. Some previously difficult items with high g_2 values therefore become easier. Items with small g_2 values, however, are about as difficult as they were prior to instruction. More detailed discussions and examples of MIRT models that accommodate differential change in item characteristics resulting from different treatments (e.g., the **c** vector in this paper) appear in Fischer (1983) and Muthén, Kao, and Burstein (1991).

of the IRT model needed to meet this goal through choices about test design, curriculum, and inferential targets.

The term $p(\Theta|\mathbf{c}_i,\mathbf{d}_i)$ represents the distribution of Θ, given the background variables $\mathbf{c}_i$, and $\mathbf{d}_i$. The proficiency variable Θ is assumed to be related to the background variables through a linear regression model, as described further below. Since a given student's item responses are not involved in determining $p(\Theta|\mathbf{c}_i,\mathbf{d}_i)$, it is referred to in this context as the prior or subpopulation distribution of Θ. The model in Equation 1 is of essentially the same form as the model used by the National Assessment of Educational Progress (NAEP), except that, for our present purposes, it is useful to distinguish the two types of background variables. The respective roles of these background variables are further discussed in the next subsection, Distinction Between Population Estimates and Aggregates of Individual Estimates.

In our presentation, we emphasize two important points.

First, unless preliminary research indicates otherwise (see the Possible Simplifications subsection), estimates of population characteristics, such as means, variances, and percentages of students at or above a particular achievement level, should be obtained directly from the model in Equation 1, rather than by aggregating the optimal student-level estimates that will also be obtained. The reason is that measurement error causes distributions of estimates that are optimal for individual students to depart from population distributions in ways that depend on the particulars of test forms. These estimation errors are particularly problematic when the characteristics of the test, as well as the nature of curricular exposure, change over assessment occasions.

Second, the model used for estimation need not be the model used for reporting. The initial scaling process serves to order the items along one or more proficiency dimensions and provides a means for linking multiple assessment forms. After the scaling process is complete, results can be transformed in various ways for reporting purposes. The possible transformations include assigning a desired range, mean, or standard deviation or reweighting the results from various content areas to correspond to the importance accorded to them in a framework

document. We recommend that reporting be based on a score scale defined by a reference test, such as a market-basket reporting scale (see Mislevy, 2003), which expresses the results in terms of expected performance on a particular set of tasks.

These points are explained in detail below.

Distinction Between Population Estimates and Aggregates of Individual Estimates

Because the Race to the Top assessment program is intended to produce both individual proficiency estimates and estimates of population characteristics, it might be assumed that estimates of population characteristics can best be obtained by aggregating the individual estimates, as suggested in the PARCC proposal (p. 63). However, both theoretical and empirical evidence show that this approach is likely to lead to inferential errors except in the case of very long and precise tests. This problem occurs regardless of the type of test score that is used, but is described here in the context of (unidimensional) item response theory models. If maximum likelihood estimates (MLEs) of individual proficiency (θ) are obtained for each student, the variance among the estimates will overestimate the population variance. If Bayesian estimates (also called expected a posteriori, or EAP, estimates) of θ are obtained, the variance among the estimates will underestimate the population variance (see Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992). Because it will lead to incorrect inferences about the spread of the students' proficiencies, the aggregation of individual θ estimates will generally lead to incorrect inferences about the percentage of students above a certain achievement level, which is one of the Race to the Top goals. Moreover, changes in test forms or differences in test form composition cause the distributions of individual estimates to change as well. In particular, these incidental modifications can affect inferences involving the tails of the distribution (such as estimates of the percentage of students who are proficient or above) by substantial amounts.

Therefore, our recommended approach is to obtain group estimates directly from Equation 1, as detailed below, without the intermediate step of obtaining individual proficiency estimates. It is necessary to include in the vector $\mathbf{d}_i$ the particular demographic variables used

to define the subgroup under consideration (e.g., gender if the group is girls or boys). Omission of demographic variables that are used in the definition of reporting groups will lead to biases in the estimation of population characteristics (Mislevy, 1991). Further information on estimation procedures is provided in the General Procedures for Estimating Population Characteristics section.

In the case of individual proficiency estimates, fairness dictates that demographic variables not be included in the estimation model. To include them would imply that two individuals with the same set of item responses, but different demographic characteristics could receive a different score, which is clearly unacceptable for decisions about individual students. With regard to the curricular exposure variables $\mathbf{c}_i$, our recommendation is somewhat more complex. We recommend that they be excluded when computing individual scores, but included when projecting individual students' future performance, which is discussed in the Predicting Students' Scores subsection later in the paper. Therefore (with the exception of predictions of future scores), individual estimates of proficiency can be obtained as the mean or the mode of the posterior distribution,

$$p(\Theta \mid \mathbf{x}_i) \propto P(\mathbf{x}_i \mid \Theta)\, p(\Theta), \qquad (2)$$

where $p(\Theta)$ is either a noninformative distribution or the overall population distribution. Under Equation 2, all students with the same observed performance would receive the same scores, which is generally desirable under considerations of fairness for high-stakes inferences about individuals.

The distribution of these per-student estimates will not generally be identical to optimal estimates of population characteristics, such as proportions of students at given proficiency levels or demographic group performance differences. This paradox results from the differential impact of measurement error on different kinds of inferences. As noted earlier, aggregating the most precise individual estimates that can be obtained from the model will, in general, yield

incorrect estimates of group characteristics (regardless of whether background variables are incorporated in the model).

Distinction Between Estimation Model and Reporting Model

Because the Θ metric is completely arbitrary, reporting results in terms of a test score is likely to be more intuitively appealing than reporting results in terms of Θ. However, since test forms will differ within and across years, it may be most useful to report results in terms of projected performance on a specially selected set, or market-basket, of tasks. (Of course, this does not preclude reporting individual item results as well.) Provided that item parameter estimates exist that allow us to link performance on these market-basket tasks to the unobserved proficiency variable Θ, we can use the observed responses $\mathbf{x}_i$ to estimate, for a particular student,

$$P(\mathbf{y}_i \mid \mathbf{x}_i) = \int P(\mathbf{y}_i \mid \Theta) p(\Theta \mid \mathbf{x}_i) d\Theta \,, \qquad (3)$$

where $\mathbf{y}_i$ represents the item responses on the reference test. These item responses can then be combined in any way that is desired to obtain the final test score (see the Point Estimation for an Individual Student on a Single Occasion subsection). For example, a decision could be made to assign greater weight to the items pertaining to the more important parts of the curriculum, or to subject matter presented later in the school year. Projecting the results from various forms of the TCSAs onto the market-basket scale provides a means of linking the forms to each other.

In the next section, we provide an example in which students' proficiency levels are reported in terms of their projected performance on the entire set of items in the domain, of which each student has taken only a subset. In the General Procedures for Estimating Population Characteristics section, we address the estimation of score distributions for population groups.

Illustration of Individual Student Proficiency Estimation

For purposes of this example, suppose we are assessing Grade 3 math, which comprises three subareas: numerical operations (N), algebra (A), and geometry (G). We assume all students taking the exams are exposed to the material in the same order—N, then A, then G—and all through-course assessments are administered at one of four fixed time points: Time 0 (before any of the three topics have been presented), Time 1 (after N), Time 2 (after A), and Time 3 (after G). The presumed distribution of items across subscales and difficulty levels is given in Table 1. The table shows, for example, that TCSA 0 consists of 10 elementary items in each of the three subscales, while TCSA 1, which occurs after numerical operations are taught, consists of a mix of elementary, intermediate, and challenging N items, along with elementary A and G items. (Elementary items are administered in the topics that have not been presented in order to provide baseline information, allowing for the measurement of growth.) We assume that test scores are to be expressed in terms of responses to the collection of 120 items that actually appear in the TCSAs (see Table 1) and that the goal is to score the test in such a way that N receives a weight of 20% and A and G each receive a weight of 40%.

Table 1. Distribution of 120 Items Across Difficulty Categories and Subscales for the Four Math Through-Course Summative Assessments (TCSAs)

Math subscale	TCSA 0			TCSA 1			TCSA 2			TCSA 3			Total
	E	I	C	E	I	C	E	I	C	E	I	C	
Numerical operations	10			2	5	3	2	6	2	2	6	2	40
Algebra	10			10			2	5	3	2	6	2	40
Geometry	10			10			10			2	5	3	40
Total	30			22	5	3	14	11	5	6	17	7	120

Note. E = elementary, I = intermediate, C = challenging. In this simplified example, it is assumed that, for each TCSA, all students receive the same set of 30 items and that no items appear in more than one TCSA. The 120 items represented in the table are assumed to constitute the mathematics domain.

In this simplified example, we assume that Θ is multivariate, consisting of three correlated dimensions, and that each item is related to ("loads on") only one of these dimensions. We assume all items are dichotomous and can be modeled using the (unidimensional) three-parameter logistic (3PL) model, estimating parameters for each subscale separately. This implies that $P(\mathbf{x}_i \mid \Theta)$ in Equations 1 and 2 can be obtained using the 3PL function.

Even within this simplified situation, we still need a model that is flexible enough to accommodate multiple dimensions and to allow for variation across the school year in curricular exposure. We can use the model from Equation 1, except that, for obtaining individual student estimates, we omit $\mathbf{c}_i$ and $\mathbf{d}_i$ from the model, as explained above.

The reporting metric for performance can take the form of a weighted sum of item scores across the entire domain (as in Bock, 1993), specifically

$$S = \sum_{j=1}^{J} w_j x_j$$

where *J* = 120 is the number of items in the domain. Note that reporting results in terms of *S* provides a metric that lends itself to studying growth or comparing performances across test forms with different compositions, even though the domain is factorially complex. Note, however, that it is probably not the ideal metric for making inferences about the individual subdomains, for example, algebra.

In this example, the weights w_j will be assigned so that $\frac{1}{J}\sum_{j\in N} w_j = .2$, $\frac{1}{J}\sum_{j\in A} w_j = .4$, and $\frac{1}{J}\sum_{j\in G} w_j = .4$, where N, A, and G denote the numerical operations, algebra, and geometry subscales. Because the domain contains 40 items in each subscale, the desired weights are .2(120)/40 = .6 for N and .4(120)/40 = 1.2 for A and G. The score S will then have a maximum of 120, the number of items. (The minimum will be the sum across the 120 items of the estimated guessing parameters from the 3PL model. In the absence of guessing, the minimum would be 0.)

Point Estimation for an Individual Student on a Single Occasion

For a particular TCSA, student *i* is administered only a subset of the items, the responses to which are denoted by $\mathbf{x}_{i,obs}$. Student *i*'s reported score on this scale, S_i^*, is the expectation of the weighted sum of responses over all items, both those administered and those not presented. In the case of dichotomous items, this score can be expressed as

$$S_i^* = E\left[S_i \middle| \mathbf{x}_{i,obs}\right] = E\left[\sum_j w_j x_j \middle| \mathbf{x}_{i,obs}\right]$$
$$= \sum_j a_j w_j x_{ij} + \int \sum_j \left(1 - a_j\right) w_j P\left(x_j \middle| \Theta\right) p\left(\Theta \middle| \mathbf{x}_{i,obs}\right) d\Theta, \qquad (4)$$

where a_j is an indicator such that $a_j = 1$ if students were administered item *j* at a given administration, and $a_j = 0$ otherwise. (As noted, we assume in this example that for a given TCSA, all students receive the same items, although this requirement is not necessary in general.) The term to the left of the plus sign in Equation 4 is the weighted sum of the observed responses on the presented items, and the term to the right is the expected sum over items not administered, given the observed performance on the administered items. In TCSA 0, for example, only 60 elementary items (20 in each subscale) will have a_j values of 1 in Equation 4 (see Table 1). All numerical operations items (administered or not) will have w_j values of .6 and all algebra and geometry items will have w_j values of 1.2. In this example, the scale in Equation 4 provides a common metric for linking all four TCSAs. Recall that although the vectors $\mathbf{c}_i$ and $\mathbf{d}_i$ do not appear in Equation 4, they would need to be included if a group characteristic were being estimated. This point is discussed further in the General Procedures for Estimating Population Characteristics section later in the paper. Equation 4 can easily be modified to accommodate items that are scored on an ordinal scale.

Although Equation 4 can be used for scoring all four TCSAs, the data entered into the model will vary across TCSAs in several ways. Of course, the vector $\mathbf{x}_{i,obs}$, containing the observed item responses will vary, as will the values of a_j, which indicate which items have

been given. Also, the distribution $p(\Theta \mid \mathbf{x}_{i,obs})$ will change from one TCSA to the next, since $\mathbf{x}_{i,obs}$ will be different.

It is important to note that a realistic testing situation is likely to necessitate a model more complicated than the one described above. Items may be scored on an ordered scale, rather than dichotomously, requiring the use of a partial credit model or graded response model. If the items are scored by human raters, the model may need to incorporate rater effects as well. The model may also need to be adapted to allow for item dependencies, assumed to be absent in typical item response models. For security reasons, students may receive different sets of items within the same test administration. Finally, whereas our example assumed a simple structure in which each item depended on only a single dimension of Θ, items will typically depend on multiple dimensions, requiring the use of a multidimensional item response model (e.g., see Reckase, 2009). In fact, the actual number of dimensions may be greater than anticipated because of overlapping capabilities required among items in different curricular segments and differences among items as to instructional sensitivity.

Predicting Students' Scores on Future Through-Course Summative Assessments

Another possible use of our model is to predict students' future performance. For example, we may want to predict a student's score on the final TCSA of the year after he or she has completed only one TCSA. For this purpose, we use a modification of Equation 4 that includes the vector $\mathbf{c}_i$. In fact, the equation for prediction includes two different sets of curricular variables: $\mathbf{c}_i$, which encodes the student's curricular exposure at the first TCSA, and $\mathbf{c}_i$*, which reflects exposure to the entire year's curriculum, corresponding to the (future) end-of-year test. The student's predicted score PS_i^* in this situation can be expressed as

$$PS_i^* = E\left[\left(S_i^* \middle| \mathbf{c}_i^*\right) \middle| \mathbf{x}_{i,obs}, \mathbf{c}_i\right] = E\left[\left(\sum_j w_j x_j^* \middle| \mathbf{c}_i^*\right) \middle| \mathbf{x}_{i,obs}, \mathbf{c}_i\right]$$
$$= \int_{\Theta}\int_{\Theta^*} \sum_j \left(w_j P\left(x_j^* \middle| \Theta^*, \mathbf{c}_i^*\right)\right) p\left(\Theta^* \middle| \Theta, \mathbf{c}_i^*, \mathbf{c}_i\right) p\left(\Theta \middle| \mathbf{x}_{i,obs}, \mathbf{c}_i\right) d\Theta^* d\Theta, \quad (5)$$

where Θ^* represents proficiency and x_j* represents item responses at the final testing occasion. The term $p\left(\Theta^* \middle| \Theta, \mathbf{c}_i^*, \mathbf{c}_i\right)$ represents the distribution of proficiency at the end of the year, given proficiency at the earlier TCSA and given the previous and predicted curricular exposure. In the absence of observed item data for the final TCSA, the inclusion of $\mathbf{c}_i^*$ especially to the extent that it differs from $\mathbf{c}_i$ is expected to substantially improve the quality of the prediction.

Modeling Growth Across Occasions

As we noted earlier, modeling growth across occasions on a single scale makes little sense when learning at one occasion is, for example, mainly in algebra, at another occasion mainly in numerical operations, and at a third occasion, mainly in geometry. Some kind of common metric is necessary to define *growth*. The discussion above includes two metrics for modeling growth in a system of TCSAs.

The first is the Θ space itself. If Θ has been defined to encompass enough dimensions to allow for item parameter invariance to hold approximately, despite differing patterns of curricula and instructional sensitivity, then change and growth can be meaningfully defined in terms of Θ. The patterns of change and growth may be difficult to communicate and interpret, however. For example, results might indicate that after completion of the algebra unit, Johnny improved by *x* in algebra and by small amounts *y* and *z* in numerical operations and geometry. In other words, growth would be expressed in terms of multiple correlated dimensions.

The second method for defining growth is the projection to a market-basket of items. Assuming that the selected model fits the data, obtaining a market-basket score for each TCSA allows the possibility of measuring academic-year growth simply by taking the difference

between the score on TCSA 4 and the score on TCSA 0. This metric is well-defined and easy to understand, and may be educationally meaningful as a weighted average of valued skills. It has the same interpretation across time points and curricular orders. It is factorially complex, however, and trends over time depend on the choices of weights and curricular orders. In other words, apparent patterns of proficiency change depend in part on incidental factors. (Note also that growth measurements based on difference scores, regardless of how they are calculated, are likely to be unreliable at the individual student level because of the high correlation between the two measurements that are being compared. See Thorndike & Hagen, 1977, pp. 98-100 for a useful discussion.)

We advise caution concerning a seemingly attractive third option for constructing a common metric; namely, defining a reporting scale by fitting a single unidimensional item response model across all testing occasions (calibrated on students who have taken all segments of the curriculum). The reason we do not recommend such an approach is that item difficulties can be expected to vary substantially relative to one another at different time points and under different curricular orders. This effect is, in fact, encouraged by the exhortation for tests to provide useful feedback to teachers during the year. This directive implies that items should be sensitive to instruction.

In the context of a unidimensional IRT model, the likely variations in relative difficulty would violate the assumptions of item parameter invariance and local independence that are necessary for obtaining well-defined individual and population estimates and for linking test forms. Experience shows that even changing the position of items can produce violations of item parameter invariance, which, in turn, can substantially affect estimated proficiency distributions. This was the case in the NAEP reading anomaly, in which the performance of 9- and 17-year olds showed an unexpected decrease between 1984 and 1986. The shapes, as well as the location, of the estimated 1986 proficiency distributions were affected (Zwick, 1991), resulting in the distortion of such statistics as the standard deviation and the proportion of students exceeding a particular scale point.

How can we avoid a model misfit problem of this kind in the proposed approach? Our intention is to use a model that is complex enough to allow the conditional independence assumption of item response theory to hold. Hence, our proposed model is multidimensional (with the structure to be determined through empirical study), allowing for a more detailed representation of proficiency.

Combining Scores Across Occasions

The definition of a market-basket metric provides a common scale for combining results across occasions. Exactly how this combination is to be accomplished is not determined by psychometric considerations alone. The question would be straightforward if there were no change at all: A simple average would provide the best estimate of the mean. But if there is growth, we might reject the use of a simple average on the grounds that it underestimates what a student has attained by the end of the year. We might wish instead to weight later performance more heavily, or use a growth model or a prediction model (see the Predicting Students' Scores subsection) in conjunction with our psychometric model to estimate final attainment, using the results from all the TCSAs. Yet another possibility is to derive a composite based on the student's best performance on each curricular unit. In any case, the decision on combining results must involve both psychometricians and subject-matter experts, and must be based on a clear conception of what the summary score is intended to measure.

General Procedures for Estimating Population Characteristics

Although Equation 4 is appropriate for obtaining individual TCSA scores, aggregating these individual scores would not, in general, be an appropriate means of estimating the score distribution for a population. In particular, the distribution of estimates obtained by taking the mean or mode of the posterior distribution of Θ for each individual (as in Equations 2 and 4) will lead to underestimation of the variance of the proficiency distribution, as noted earlier. Estimation of the distribution of scores like those defined in Equation 4 for a population of interest can be achieved using direct estimation solutions (e.g., Cohen & Jiang, 1999; Mislevy,

1985) or multiple imputation procedures paralleling those used in NAEP. As detailed below, the latter approach involves estimating the posterior distribution of Θ for each individual and then drawing a value (an imputation or plausible value) for use in computing the statistic of interest. A total of *M* draws from each individual's distribution is taken, creating *M* datasets. Now suppose the statistic of interest is the percentage of students exceeding a particular scale value and suppose *M* = 5, as in NAEP. The percentage is recomputed for each of the five datasets, and the average (say, *P*) of the five results is used as the final estimate. The variation among the five initial results provides an estimate of measurement error, which is later added to the sampling variance to obtain the (squared) standard error of *P*. Related discussions are given by Mislevy (2003, pp. 39–41), Johnson and Jenkins (2005, pp. 8–12), and Thomas (2000, p. 355).

The following steps outline the procedures for estimating characteristics of a population proficiency distribution in terms of projected performance on a set of market-basket items.

1. Fit the chosen IRT model to all items to be used in the TCSAs for a given year, as well as any additional items to be used as a basis for reporting (i.e., the market-basket items, which may consist of a targeted selection of items from earlier years). After the scales have been established in a start-up year, the calibration phase serves as an opportunity to introduce new items into the scale. Treat all estimated item parameters $\hat{\boldsymbol{\beta}}$ as known for the remaining steps. Ideally, the samples used for item parameter estimation (calibration) should include students who have had varying degrees of exposure to the material in question. Using data from only a few occasions or curricular patterns–for example only data from the end of the year—could hide shifts in item response patterns that occur during the year because of differences in curricular order and instructional sensitivity. Systematic shifts in difficulty due to instruction could then manifest themselves as unaccounted-for model misfit, causing distortions in the estimated proficiency distributions, as mentioned in the Modeling Growth Across Occasions subsection.

2. Using a latent variable regression model for $p(\Theta|\mathbf{c},\mathbf{d},\boldsymbol{\Gamma})$, the conditional distribution of Θ given background variables **c** and **d** (Mislevy, 1985), obtain the posterior distribution of the parameters $\boldsymbol{\Gamma}$ of the regression model, namely $f(\boldsymbol{\Gamma} \mid \mathbf{x},\mathbf{c},\mathbf{d},\hat{\boldsymbol{\beta}})$. (These are subpopulation distributions that must be used to construct imputations for Θ so that they will echo back consistent estimates of statistics involving **c** and **d**.)

   Repeat Steps 3-7 below for each of the *M* datasets:

3. Draw a value for each regression parameter from $f(\boldsymbol{\Gamma} \mid \mathbf{x},\mathbf{c},\mathbf{d},\hat{\boldsymbol{\beta}})$. Treat these values $\tilde{\boldsymbol{\Gamma}}$ as known in the remaining steps.

4. For each sampled student, compute the posterior distribution $p(\Theta \mid \mathbf{x}_i,\mathbf{c}_i,\mathbf{d}_i,\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\Gamma}})$ as in Equation 1. (For simplicity of notation, dependence on estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ was not explicitly noted in Equations 1-4.)

5. For each student, draw a value from this distribution, denoted here as $\tilde{\Theta}_i$.

6. For each of the market-basket items, use $\tilde{\Theta}_i$ from Step 5 to draw a value from $P(\mathbf{y}_i \mid \tilde{\Theta}_i)$, producing a set of imputed item responses, $\tilde{\mathbf{y}}_i$.

7. Obtain $S(\tilde{\mathbf{y}}_i)$, the desired function of those imputed responses, such as a simple sum or weighted combination.

8. Estimate the desired population characteristic from each of the *M* data sets and then average the results, as described above.

An advantage of the multiple imputation approach outlined above is that it can be combined with whatever methodologies are needed to deal with the sampling variance of students (Mislevy, 1991; Rubin, 1987). NAEP, with its multilevel sampling design and weighting scheme, uses jackknife procedures for this purpose (see Johnson & Rust, 1992, for a general description and National Assessment of Educational Progress, 2011, for updated information). A

simple random sample would use familiar standard errors of the mean for the sampling component of variance for group characteristics. In the present context, an improved estimate could be obtained by taking into account the clustering of students within schools and districts. The computation of sampling errors in census-type studies is intended to provide an indication of variation that a user should allow for, in light of expected year-to-year variation in student populations.

Discussion and Recommendations

We have outlined a model for use in analyzing data obtained from TCSAs administered as part of the Race to the Top Program. The analysis model is similar to the one that has been used by NAEP for the last 25 years. Similar models and analysis procedures, involving multiple imputations, are also used by TIMSS and PISA (see von Davier, Sinharay, Oranje, & Beaton, 2006). An advantage of applying an analysis model similar to that used in other large-scale assessments is that there is a body of applicable research that users can draw upon (e.g., Johnson & Jenkins, 2005; Thomas, 1993, 2000). Our discussion extends the practices of NAEP, TIMSS, and PISA in that we propose reporting results in terms of a market-basket of items, a technique that is intuitively appealing and allows for simple approaches to the measurement of growth.

A disadvantage of the proposed model is its complexity. Understanding and implementing the model requires a certain amount of statistical sophistication. In addition, the model cannot be estimated with off-the-shelf software, and troubleshooting can be less than straightforward. Why choose such a complicated model for this application? In short, the answer is that the fewer the constraints on the testing situation, the greater the demands on the analysis model. As we noted, the Race to the Top testing situation involves multidimensional subject matter, multiple test forms, complex item types, and varying patterns of instruction. The Race to the Top program seeks to estimate growth for individuals, provide estimates of subtle characteristics of population and subpopulation distributions, gather data on different aspects of a domain at different time points, and give teachers feedback using instructionally sensitive items. A simple analysis model cannot accommodate all these features.

It is worth noting that the limitations of a test analysis model often do not become apparent until the second time a test is given—the first set of results may look entirely reasonable. An implausible change in student performance between Time 1 and Time 2 may provide the first indication that the analysis model is inappropriate—for example, it may not be elaborate enough to accommodate changes in test forms, instructional exposure, and other features of the testing situation. The demands on the model are further exacerbated in a high-stakes environment in which results will be heavily scrutinized and used to compare students, classrooms, schools, districts, and states.

Possible Simplifications

Can the complexities of the general model be reduced in practical applications? This section offers some thoughts on this question. As detailed further in the Recommendations subsection in this paper, explorations with pilot data will help determine whether these simplifications are feasible. An advantage of having an overarching framework, such as the one presented above, is that streamlined approaches can be compared against more elaborate procedures in terms of their effects on particular inferences. Informed decisions can then be made regarding operational procedures. Three potential simplifications that might be considered are using special booklets administered to samples of students for certain inferences, simplifying the IRT model, and simplifying the population model. We briefly discuss each of these in turn.

Using special booklets administered to student samples. The complexity of the full model in Equations 1-4 arises in part from a desire to avoid constraining test form designs. For example, complex item types, including extended performance items, are of interest in Race to the Top assessments. One approach to this situation would be to use relatively unconstrained test forms for certain purposes, such as informing classroom instruction, while using more constrained test forms, administered to only a sample of students, for other inferences, such as comparisons of schools, districts, and states (Bejar & Graf, 2010). These more rigorously constructed forms, which could consist of parallel tests with controlled mixtures of machine-

scoreable item types, could be seeded into test administrations. Basing the primary reporting scale on only the more constrained forms would allow more flexibly created forms to be used for instructional purposes, and could also facilitate the use of a simpler IRT model for the primary reporting scale, as described in the next section.

Simplifying the IRT model. The notice inviting applications (Department of Education, 2010) allows for gathering responses to instructionally sensitive items administered in different mixes at different times under different curricular orders. The most conservative approach to analyzing these responses with IRT is to anticipate a MIRT model with items loading on multiple dimensions, possibly in complex patterns.[4] However, with appropriate attention to test construction, a simpler model that assumes a particular multidimensional structure could suffice. Simpler models include joint unidimensional scales, as in the example of the Illustration of Individual Student Proficiency Examination section, a bi-factor model (Gibbons & Hedeker, 1992), or a Saltus-like model (see Mislevy & Wilson, 1996; Wilson, 1989).

Simplifying the population model. A question that is likely to occur to the Race to the Top consortia is whether a population model per se is needed at all. Could population characteristics be inferred from the aggregate of the individual proficiency estimates? As we noted earlier, this would be a workable strategy in the case of a test with consistently high reliability; that is, the reliability would not only need to be high, but would need to be roughly constant over assessment forms and occasions. Fluctuations in reliability from one assessment to the next have the potential to distort inferences about growth, particularly if they involve the tails of the distribution (e.g., the change in the percentage of students who are proficient or above). It is possible that the approach outlined in the Using Special Booklets subsection, if implemented with a highly reliable test, could allow the possibility of estimating population characteristics by aggregating individual estimates for the students receiving the special

[4] In a MIRT model, shifts in item difficulties associated with different curricular patterns and instructional sensitivities can be captured as differences in proficiency profiles rather than as potentially distorting sources of misfit in a simpler IRT model. See Footnote 3.

booklets. This, again, could be investigated in the pilot and field test phases specified in the notice inviting applications (Department Education, 2010). It is worth noting that implementing a simplification of this kind would not only eliminate the need for separate estimation procedures for individual and group characteristics, but would increase the likelihood that scores could be reported quickly and economically.

Recommendations

This section contains our general recommendations about psychometric strategies for TCSAs.

Recommendation 1. Recognize the tradeoffs between inferential demands and procedural simplicity. The more demands that are made of the scaling and reporting model—that it accommodate complex items of varying instructional sensitivity, for example—the more complex the model needs to be. As demands are reduced, simpler approaches become more feasible.

Recommendation 2. Take advantage of the pilot and field test periods to evaluate psychometric approaches. For example, tests of IRT model fit can help to determine whether including complex tasks in the summative assessment scale is feasible. Pilot investigations can serve to determine if the IRT and population models can be simplified, as we note in the Possible Simplifications subsection. Pilot testing can reveal whether it is possible to relax the claims for the assessment system or add constraints to the curriculum or the assessment designs so that simpler models or approximations will suffice.

Pilot testing should include the collection of response data from students who are at different points in the curriculum and who have studied the material in different orders. This data collection would allow exploration of the dimensionality of the data with respect to the time and curricular exposure variables that must be accommodated in the TCSA paradigm. Only by examining data of this sort can we learn whether simpler IRT models can be employed. Estimation of parameters for extended response tasks, including rater effects, should be studied in pilot testing as well, since these items tend to be unstable and difficult to calibrate into existing scales. How well will they work in the anticipated system?

A data collection of this kind would also support explorations of the estimation of the posterior distribution of proficiency, $p(\Theta \mid \mathbf{c}_i, \mathbf{d}_i, \boldsymbol{\Gamma})$. How much data is needed for stable estimation? Are effects for $\mathbf{c}_i$ small enough to ignore? Again, data collection at a single occasion will not be sufficient to investigate these issues.

Finally, pilot testing should gather some longitudinal data from at least a subsample of students for purposes of studying growth modeling and combining results over occasions. Little is known about either the stability or the interpretability of results in this context.

Recommendation 3. For any assessments used to make comparisons across schools, districts, or states, recognize the importance of establishing and rigorously enforcing shared assessment policies and procedures. The units to be compared must establish policies concerning testing accommodations and exclusions for English language learners and students with disabilities, test preparation, and test security, as well as rules concerning the timing and conditions for test administration (see Zwick, 2010). Careful attention to data analyses and application of sophisticated psychometric models will be a wasted effort if these factors are not adequately controlled.

In summary, the TCSAs proposed as part of the Race to the Top initiative are expected to satisfy a multitude of inferential demands and, correspondingly, present a host of psychometric challenges. While the adoption of a simple method for scaling and linking the TCSAs may be appealing, it could also lead to incorrect inferences about student proficiency. Inferences about the proportion of students attaining or exceeding a certain achievement level (which involve the tails of proficiency distributions) and inferences about change over time are particularly susceptible to error. A more complex analysis model, such as the one we have outlined, offers some insulation against changes between assessment occasions in the number, format, and instructional sensitivity of the test items, as well as changes in the students' patterns of curricular exposure. We recommend that the pilot and field test periods be exploited to test out various analysis methods and to explore possible simplifications. Finally,

we recommend careful attention to assessment policies and procedures so as to maximize the validity of comparisons across students, schools, districts, and states.

References

Bejar, I. I., & Graf, E. A. (2010). Updating the duplex design for test-based accountability in the twenty-first century. *Measurement: Interdisciplinary Research & Perspectives, 8,* 110-129.

Bock, R. D. (1993). *Domain referenced reporting in large scale educational assessments*. Paper commissioned by the National Academy of Education for the Capstone Report of the NAE Technical Review Panel on State/NAEP Assessment.

Cohen, J., & Jiang, T. (1999). Comparison of partially measured latent traits across nominal subgroups. *Journal of the American Statistical Association. 94*, 1035–1044.

Department of Education. Overview information; Race to the Top Fund Assessment Program; Notice inviting applications for new awards for fiscal year (FY) 2010. 75 Fed. Reg., 18171-18185 . (April 9, 2010).

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3–26.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item. bi-factor analysis. *Psychometrika, 57,* 423–436.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175-190.

Johnson, M. S., & Jenkins. F. (2005). *A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress* (ETS Research Report No. RR-04-38). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (2003). *Evidentiary relationships among data-gathering methods and reporting scales in surveys of educational achievement* (CSE Technical Report No. 595). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993-997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133-161.

Mislevy, R. J., & Wilson, M. R. (1996) Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61,* 41-71.

Muthén, B., Kao, C.-F., & Burstein, L. (1991). Instructional sensitivity in mathematics achievement test items: Applications of a new IRT-based detection technique. *Journal of Educational Measurement, 28,* 1–22.

National Assessment of Educational Progress. (2011). *NAEP weighing procedures—Replicate variance estimation for the 2007 assessment.* Retrieved from http://nces.ed.gov/nationsreportcard/tdw/weighting/2007/weighting_2007_repwts_appdx.asp

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

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley.

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

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2,* 309-322.

Thomas, N. (2000). Assessing model sensitivity of the imputation methods used in the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics, 25*, 351-371.

Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4th ed.) New York, NY: Wiley.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. E. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics*. Amsterdam, the Netherlands: Elsevier.

Wilson, M. R. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276-289.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*, 10-16.

Zwick, R. (2010). *Measurement issues in state achievement comparisons* (ETS Research Report No. RR-10-19). Princeton, NJ: Educational Testing Service.