Combining Multiple Indicators

Lauress L. Wise, HumRRO
Paper for the PARCC Technical Advisory Group
September 6, 2011

Overview

The Partnership for Assessing College and Career Readiness (PARCC) plans to collect a wide range of information about student achievement in English/language arts (ELA) and mathematics in grades 3 through 8 and in high school. All of this information must be summarized in two key indicators. The first is a status measure, indicating the extent to which each student is proficient in the reading or mathematics knowledge and skills targeted for their current grade by the Common Core State Standards (CCSS). The second is a growth measure indicating how much the student has learned during the year. Both measures have important uses in reporting to teachers, parents, and the students themselves. Aggregated across students, the status measure will tell us how well we are doing in meeting key educational objectives. Aggregation of the growth measures will be used in assessing the value added by schools and, perhaps also by individual teachers.

Options for developing measures of growth from individual student status measures will be explored in a separate paper, but the need for measures of growth means that the status measure has to be sensitive to a wide range of student achievement levels in addition to a single proficiency cut-off. In addition, where more specific indicators are reported in addition to the overall summary indicator, it may be desirable to report growth on each of these separate indicators as well as growth on the overall achievement composite. Issues and considerations for combining separate growth indicators are essentially the same as for combining separate status indicators, so that the following discussion applies equally to both situations.

The primary focus of this paper is describing and evaluating alternatives for combining assessment results across test administrations, content areas, and item formats. The paper also addresses how test results might be combined to create additional reportable scores. For ELA, separate scores for reading and writing are planned. For mathematics, each elementary grade will have a focal content area score, in addition to an overall score. An additional concern, addressed later in the paper, is how to create overall readiness scores at the high school level using results from multiple end-of-course tests.

An earlier concern, how to combine results collected at different points of the school year, has largely been eliminated with the PARCC decision to make the first two quarterly assessments optional and not include them in overall summative measures. States and districts who do administer the earlier quarterly assessments will need to consider how to use the results to assess mastery of particular parts of the curriculum, mid-year growth, and projected end-of-year status and growth. Options for using interim assessment results were described in an earlier paper (Wise, 2011) and will be touched on briefly at the end of this paper.

Current PARCC plans call for combining results from the 3rd quarter through-course assessment (TC3) and the end-of-year assessment (EOY) into an overall summative measure. It is not currently clear whether results will be reported separately for these two assessments. TC3 will be administered at an earlier point, but it will require extensive scoring processes. The EOY assessment will be machine scored. It is likely that results from these two assessments will be available at approximately the same time, thus eliminating any advantage from reporting TC3 scores separately. Pending decisions of whether to report separate scores, for TC3 versus EOY or for different areas of knowledge and skill will have implications for the design and equating of each assessment. Issues with respect to how a combined score is derived from these two assessments are essentially the same as the issues for combining results across content and item formats discussed in the remainder of this paper.

General Alternative Approaches to Combining Indicators

There are several general approaches to creating an overall indicator of student achievement. The first approach, joint calibration, is to build an overall indicator directly, modeling the content covered by the assessment as a single underlying dimension and scaling all test questions as measures of this single dimension. A second approach is to build separate indicators of achievement as measured by items from different content areas or using different item formats and then form a weighted composite of these separate indicators. Two variations of the second approach are described below, one in which the relative weight given to each of the separate indicators is based on policy judgments of their relative importance and the other in which component weights are empirically derived to maximize prediction of some criterion or to maximize internal consistency estimates of score reliability. A final approach considered here is a multiple hurdles approach where students must pass separate proficiency cut-offs on some or all components of the test to be considered successful in mastering the overall content covered by the assessment. Details of each approach are discussed next.

Joint Calibration

Perhaps the most common approach to combining results across content areas and item formats is joint calibration. This approach involves fitting a single, unidimensional IRT model to responses to all of the items, regardless of content or format. To be sure, there is wide variation in the IRT models chosen, particularly for polytomously scored items (e.g., Samejema, 1969; Masters, 1982; and Muraki, 1992), but the basic assumption of unidimensioinality is the same regardless of the specific IRT model chosen. Joint calibration is used in most states and, increasingly, in developing NAEP scores (NCES, 2009).

Policy-Weighted Composites

A second approach to combining assessment results from different content areas and/or item formats is to form a weighted composite of the separate pieces, using weights defined by

policy judgments. Leucht and Miller (1992) suggest dividing up a multi-dimensional domain into separate unidimensional measures. The question of how to put these separate measures into an overall indicator is most often resolved by falling back on policy judgments as to their relative importance. Nearly all state assessments are developed from blueprints that specify the number of score points to be assigned to each area of content. Unfortunately, all too often the number of assigned score points is based on the number of specific standards or objectives that have been adopted for each content area rather than any direct judgment of the importance of the content area. A key question in adopting policy-derived weights is whether coverage should be spread broadly across a wide domain, as is often the case when weights are based on the number of objectives, or whether there are a few areas of content that are core and significantly more important than other areas.

An additional question for developing policy-weights is how to handle underlying dimensions that are defined more by item formats than by content. In some cases, such as the California High School Exit Examination, extended constructed response items cover unique content (writing application) as well as having a unique item format. For PARCC, however, extended constructed response or performance tasks are likely to be targeted for a number of different areas of the CCSS. Empirical factor analyses often show that items with different formats load on different dimensions, even when CR and MC items are intended to measure the same content standards. One potential resolution of this problem is to target CR and MC items for different standards or objectives. After all, if MC and CR items are thought to measure mastery of the same standard, we might not need the more costly CR items in the first place. In any event, a description of the specific skill or cognitive processes measured by each item format would be needed before policy judgments about the relative importance of the factors related to item format can be gathered.

In educational assessments, it is common to allocate total score points across content and skill areas as a reflection of policy judgments about the importance of each area. An alternative approach, more common in psychological testing, is to standardize each component score to have a common variance (most often 1.0) and then apply policy-derived weights to the standardized component scores. In this alternative approach, the policy weights reflect each components contribution to the variance of the resulting composite scores. Similarly the weights determine the contribution of each component to covariances between the overall indicator and external variables of interest. In practice, the distributions of items used to measure each content area are similar enough that the variance of each component is roughly proportional to the number of score points. In this case, the two approaches to applying policy weights are nearly equivalent.

Empirically-Weighted Composite

An alternative to weights based on policy judgments is to estimate weights that maximize the prediction of one or more important outcomes. Commonly used in developing employment or job performance tests, this approach has rarely been applied to educational assessments, due largely to the unavailability of criterion measures and/or the lack of agreement regarding outcomes the achievement measures should be expected to predict.

NAEP provides two interesting examples of the promise and perils of predictive weights (Kifer, 2001; Jones and Olkin, 2004). When achievement levels were initially developed, the original policy descriptions of the levels talked about being prepared for work at the next level. The National Assessment Governing Board (NAGB) quickly realized, however, that they could not provide empirical support for assertions about preparedness and so changed the descriptions to simply talk about mastery over challenging content (Stufflebeam, Jaeger, & Scriven, 1991). Now, fast forward 30 years. NAGB is now engaged in a significant effort to be able to interpret 12th grade NAEP results as indicators of preparedness for college and careers. NAGB asked Achieve to suggest changes to content frameworks for reading and mathematics that would improve their use as measures of preparedness. Empirical data on college outcomes and judgments about college and career preparedness are being collected to support preparedness interpretations of NAEP reading and mathematics scores, but there are no current plans to use predictive weights either to combine reading and mathematics scores into an overall indicator of preparedness or to combine components of reading or components of mathematics scores to improve their prediction of key outcomes.

The CCSS were designed to lead up to preparedness for college and careers at the end of high school. This design strongly suggests the need for empirical data on the extent to which scores from the end of high school predict that students will not need to take remedial courses in the first year of college, will achieve successful grade point averages in their course work, or will succeed in job training programs and begin to progress in their chosen careers. Assessing college and career readiness is a key goal for PARCC.

The recent development of longitudinal student data bases in many states now makes it feasible to predict success from grade to grade, leading up to college and career readiness at the end of high school. The question of what students need to know and be able to do to be successful in the next higher grade is open to empirical evidence. The relative importance of success in different components of the current-year curriculum for predicting success in subsequent years could be used to determine effective weights for combining the components into an overall indicator that is maximally predictive of future success.

While empirically derived component weights have some appeal, they are unlikely to be used right away. At the very least, it will take several years of collecting and analyzing linked achievement data before stable and defensible empirical weights could be developed. In the meantime, it will be necessary to use either joint calibration or policy-derived component weights to create overall summative measures.

An alternative and more immediate possibility for the empirical development of component weights is to identify composites that maximize score reliability as measured by internal consistency indicators. This approach typically involves the use of two-level factor analyses. Separate component scores are estimated from discrete sets of item responses at the first level. Then weights are derived for estimating the linear composite of the component scores that best explains the correlations among the components. See Figure 1 for an illustration of this approach. In this approach, components that were highly correlated with other components would have higher weights than components that were largely independent of the other components.

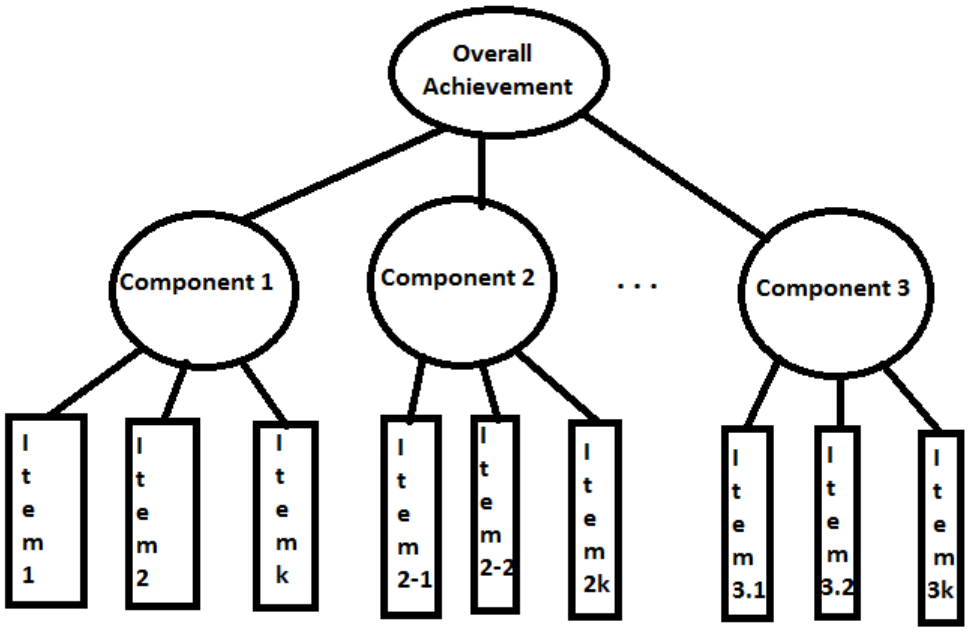

Figure 1. Two-level factor analysis model for overall and component scores

It is likely that the two-stage factor model would yield overall scores very similar to those generated through joint calibration. If the component scores are very highly correlated, joint calibration would generate item loadings on the general factor that would be essentially the same as the loadings on the separate component factors. Even if the components are only moderated correlated, the underlying dimension in a joint calibration will approximate the first principle component underlying the item correlations (Drasgow and Parsons, 1983).

Multiple Hurdles

Each of the above approaches assumes a compensatory model. Higher performance on some groups of items or components will make up for lower performance on others. An alternative is a multiple hurdles approach where successful performance on all or some number of the components is required for overall success. This approach is often used with high school graduation tests. In California, students must pass both the ELA examination and the mathematics examination to qualify for a diploma. In other states, students must pass some minimum number of end-of-course tests to receive a diploma. Multiple hurdle models might provide a useful alternative for combining different high school end-of-course tests into an

overall indication of readiness for college and careers. The judgments required to support a multiple hurdles model may be contentious and difficult to obtain.

Strengths and Weaknesses of Each Approach

Each of the approaches to combining multiple indicators described above offers some advantages but also has some disadvantages. Some of the major potential strengths and weaknesses of each approach are described briefly here. In the following section, research is suggested that would better elucidate advantages and issues in implementing each approach.

Joint Calibration

Potential Strengths. A clear advantage of joint calibration is that neither policy judgments nor external criteria are required. A second possible advantage is that item calibration, equating, and development of scoring tables used to generate reported scores can all be done at the same time.

Potential Weaknesses. A key disadvantage of joint calibration is that it does not support subscores. Where there is an intention to report separate scores so as to provide diagnostic information on individual students or aggregate information curricular evaluations, it may make more sense to scale the separately reportable scores first and then develop explicitweights for combining them into an overall indicator. A second potential disadvantage is that all of the data must be available before analyses can begin. Joint calibration might not work well if there was an intention to report interim results. Joint calibration is unlikely to be feasible at all as a means of combing results from different high school end-of-course tests.

Another, potentially serious disadvantage of joint calibration is that if the domain covered by the assessment is truly multidimensional, the items or item formats assessing relatively unique aspects of this domain may be given little weight as item weights tend to reflect loadings on the first principle component of the multiple dimensions (Drasgow and Parsons, 1983). From a policy perspective, it may be desirable to give items tapping unique parts of the content domain more weight (as this part of the domain is otherwise underrepresented) rather than less weight.

Policy-Weighted Composites

Potential Strengths. Judgments were used in creating the CCSS, so it is reasonable to use expert judgment to create weights to be applied to separate components of the assessment. This approach is applied simply and transparently in test blueprints that assign different maximum score points to each component. Note that weights reflecting policy judgments about the relative importance of different areas of content and skill will likely also have an impact on messages about curriculum and instruction in addition to having an impact on the resulting summary measure.

Potential Weaknesses. Experts may disagree regarding the relative importance of different components and some would argue that empirical data on predicting important criteria are still needed to supplement or validate the expert judgments.

Empirically-Weighted Composites

Potential Strengths. Predictive validity studies can provide unimpeachable evidence of the nature and level of content that students should master to be successful in the next grade or after high school. Perhaps more importantly, results from such studies show clearly why mastery of this content is important.

Potential Weaknesses. Experts may disagree as to which outcomes provide critical criteria and measures of these outcomes will be imperfect. The most convincing criteria are likely to require longitudinal data linking student test scores at one point to outcomes one or more years later. Collecting such data takes time and so interim composites will be needed while the empirical data are collected.

Multiple Hurdle Composites

Potential Strengths. Multiple hurdles composites ensure that students demonstrate adequate levels in each of several components of mastery. This approach is most appropriate when experts believe that exceptional performance in one area cannot fully compensate for sub-par performance in another area.

Potential Weaknesses. Multiple hurdle measures tend to be discrete, often dichotomous variables, creating difficulties in assessing growth at different levels of initial performance. This approach also requires judgments about the skills and. more importantly, the level of each skill required for each overall level of proficiency and experts may not agree.

Research Needs

A body of research will be needed to support decisions about the most appropriate ways to combine multiple indicators into overall measures of achievement and growth. Ideas for key research that could be conducted in conjunction with item development and tryout are sketched here. Obviously, a more complete agenda for research covering a range of additional topics will be mapped out as test development begins.

Determination of Intended and Actual Uses of Score Information

The first area of research begins with capturing policy regarding scores to be provided in addition to overall end-of-year summative scores. Current plans call for a separate score for a focal area of mathematics at each grade (3-8) and possibly separate reading and writing scores. It is not clear that any subscores are planned for high school course-based assessments, but that

needs to be verified. Teachers and administrators are likely to want subscore information, both for diagnosing individual student needs and also for evaluating strengths and weaknesses of current curriculum and instruction. A first phase of research would be analyses of the degree of consensus, across states and stakeholder groups, regarding the subscores to be provided. Where subscores are provided, the process for developing and then equating them from year to year will have to be specified. In addition, there will likely be strong demands for clarity regarding the relationship of the subscores to the overall achievement indicator. Summary scores derived through joint calibration will be difficult to explain if the overall score cannot then be derived from the subscores in some meaningful way.

A second phase of research on actual uses might be conducted in conjunction with field testing of new test items and forms. Interviews with teachers and school officials should be conducted to determine the ways in which score information might be used, intended or unintended, and the possible impact of such uses on curriculum and instruction and ultimately on student achievement.

Analyses of the Dimensionality of Item and Component Scores

As new measures are developed and piloted, it will be important to begin analyses of the dimensionality of these measures. Dimensionality may result from relevant differences across content or skill categories or unintended (criterion irrelevant) differences across item formats or time of assessment[1]. Significant relevant dimensionality will suggest the need to ensure that each dimension is appropriately represented in the overall indicator. On the other hand, if most or all of the correlation among item and component scores is explained by a single general factor, then joint calibration or empirical weights that maximize overall score reliability could be considered.

Confirmatory analyses should be used to analyze covariances among item or component scores, starting with a single general factor and then fitting models of increasing dimensionality. Research would proceed until satisfactory fit is found or until additional dimensions do not significantly improve model fit. Higher dimensional models would be based on expert judgment about how items might best be clustered into content and/or skill categories. Results for such models could be compared to results from models where the same number of dimensions was used, but the dimensions were based on potentially irrelevant factors (e.g., item format or time of administration).

Impact of the Timing of the TC3 Administration Date

Due largely to the time required for scoring, TC3 results will likely be treated as though they were end-of-year measures and may not be reported separately (except to the extent that they cover separate content that is to be reported separately). During pilot testing, it should be possible to administer some TC3 items and possibly whole forms at the very end of the year and others at the end of the third quarter as intended in operational use. This would provide a test of

[1] Note that to the extent that different item formats are intentionally linked to different content or skill differences, dimensionality across item formats may be a valid reflection of the dimensionality of the target domain.

the extent to which TC3 results might underestimate end-of-year performance. If such differences were found, it might be important to consider adjustments to TC3 results prior to combining them with EOY results or, at least, to take such differences into consideration in setting achievement level standards.

A related area of research concerns sensitivity to the timing of the TC3 assessments. There is likely to be some variation in when TC3 is administered with respect of the different school year calendars used by different states and districts. Comparison of result from relatively early and relatively late TC3 administrations will help in deciding whether adjustments for time of administration might be needed.

Pilot of Joint Calibration Procedures

If joint calibration is a possible option for at least some of the combined indicators, significant pilot testing of options for implementing this procedure will be required. Different IRT models and different procedures for identifying calibration samples should be tried out using pilot and field test data. Outcomes include the frequency of dropping items due to lack of model fit, the stability of parameter estimates, and calibration consistency across states and demographic groups.

Pilot of Judgments Regarding Policy Weights

There is not an available body of research on the process of establishing test blueprints or other ways of identifying the relative importance of different components of content and skill standards. For the most part, consensus is achieved through direct discussions among groups of experts. Policy capturing studies are more common in the psychological testing literature (Connolly and Ordonez, 2003). In this approach judges are presented with cases that have different combinations of component scores and asked to make an overall rating. Their policy is captured by estimating weights for predicting their judgments from the component scores for each case.

However policy judgments are captured, an important study would be to use independent panels of judges to develop recommended weights for different content or skill areas and then examine consistency in results across the different panels.

Empirical Prediction Studies

The most significant area for research on how to combine multiple measures of student achievement will be empirical prediction studies. The first stage of these studies would be identifying and obtaining a consensus around criteria that the achievement composites are intended to predict. Next criterion data would be collected for students who have taken each targeted assessment. The data would then be analyzed to estimate the target assessment component weights that maximize predication of the corresponding criteria.

For Grade 3 through 8 measures, the most likely criterion will be the overall achievement measure at the next higher grade. More significant work will be needed to identify appropriate criteria for the high school end-of-course measures. It is likely that multiple college-related criteria (need for remediation, first-year grades, graduation) will be needed in addition to criteria related to career progress and success. An optimal sequence would likely be to work with the high school measures first and then work backwards, grade by grade, so that empirical weights estimated for one grade could be used to generate the criterion composite for estimating weights at the next lower grade.

Impact of Timing of Optional TC1 and TC2 Assessments

Because they are currently optional, the timing administration of the first two through course assessments (TC1 and TC2) will likely vary widely in comparison to the timing of TC3. In addition, the curriculum that precedes these assessments will likely also vary considerably from state to state and district to district. During pilot and field testing of TC1 and TC2 items, it should be possible to study the effects of variation in timing and curriculum on scores for TC1 and TC2 items and also study variation in the relationship between TC1 and TC2 scores and EOY scores.

Recommendations

Recommendations are provided here in the form of sample decision trees. These trees are intended to provide a starting point for discussion of alternatives for combining multiple indicators. Refined decision trees based on further discussion could then be used in framing recommendations for consideration by PARCC policy-makers. Sample decision trees are provided for three situations: (1) combining TC3 and EOY measures into overall indicators of achievement in grades 3 through 8, (2) combining end-of-course test results into overall indicators of readiness for college and careers, and (3) uses and interpretations of optional TC1 and TC2 assessment scores.

Decision Tree 1: Combining Third Quarter and End-of-Year Assessment Results

1. **Are subscores needed?**

   Separate reading and writing scores may be needed for ELA or there may be some desire to report separate scores for the TC3 and EOY assessments. In mathematics, a separate score is planned for a focal area, but not for the rest of the content covered in each grade. Note that a decision to report separate scores for different components has significant implications for year-to-year equating designs (discussed in a separate paper). If separate scores are needed, restart the decision tree for each separate score and then use policy judgments to determine weights for each of the separate scores. Where separate scores are not needed, proceed to the next step.

2. **Are scores from different content areas or item formats essentially unidimensional?**

   There are several ways of checking item response data for dimensionality (Stout, 1987; Yen, 1984). If tests indicate a strong general factor, joint calibration should be used. If there is not a strong general factor, separate component scores should be established for each dimension and component weights should be developed based on policy judgments to ensure that each component receives appropriate weight. Joint calibration may ignore or underweight secondary factors in the achievement results. Where data are multidimensional, the next step should also be considered.

3. **Will external criteria measures judged important by experts be available?**

   If separate components are defined, whether reported or not, the policy weights applied to these components should be validated to the extent feasible. For grade 3-8 assessments, it should be feasible to determine the relative weighting of the components that best predicts overall success at the next grade, with the Grade 8 assessment predicting success on the high school readiness indicator. Working backwards, revising the 8^{th} grade composite to predict high school readiness, then revising the 7^{th} grade composite to best predict the revised 8^{th} grade composite, and so forth will most clearly establish the chain of progression to college and career readiness envisioned in the CCSS. Predictive validity studies, if feasible, can be used in reviewing and revising content specifications over time as well as in refining achievement composites for each subject and grade.

<u>Decision Tree 2: Combining High School End-of-Course Tests</u>

For the high school assessments, separate scores will obviously be needed for each course since these scores must be reported as the course is completed. Even if assumptions of unidimensionality were plausible, joint calibration would likely not be feasible because of the different times at which data from the different courses are available. Here, the question of multiple hurdles is a more serious concern, so the decision trees for building an overall readiness composite starts there.

1. **Do experts believe that minimum mastery of each course is required?**

   If the answer is yes, a multiple hurdles model should be considered. If experts believe that higher performance on some end-of-course tests can compensate for lower performance in others, then policy-weighted composites should be developed. Analyses of the dimensionality of results across different courses may help inform the decision to consider a multiple hurdles approach. If clearly separate dimensions are not identified, there would be less value to a multiple hurdles approach. In this case factor loadings might be considered as an alternative to weights derived by policy.

2. **Will external criteria measures judged important by experts be available?**

   Until external criteria are available, the relative weights assigned to different end-of-course assessment scores will necessarily be set by policy judgments. It is important that

such judgments include college educators and individuals who work in technical training as well as teachers of the high school curriculum. When readiness criterion measures have been identified and longitudinal data are available, initial policy judgments can be confirmed or modified through analyses of the contributions of different scores to the prediction of the external criteria.

Decision Tree 3: Uses and Interpretations of TC1 and TC2 Assessment Scores

1. **Do Earlier TC results provide accurate estimates of end-of-year status?**

   If earlier TC assessments cover material that is taught in earlier quarters, they may be essentially end-of-unit tests. If so, it is possible, that mastery at the end of the unit is approximately the same as end-of year mastery of this same material. In this case, TC1 and TC2 scores could be used interchangeably with EOY scores as indicators of mastery of targeted content. If not, then it is necessary to proceed to the next step.

2. **Do Earlier TC results predict EOY results in the same content area(s)?**

   Alternative prediction models might be considered. One assumes constant growth across students so that EOY status results would be a fixed increment above TC1 or TC2 status. A second model assumes constant growth across students at the same TC1 or TC2 status level. The expected increment from TC1 to TC2 to EOY would vary by initial level, but would not vary across students at the same level. A third model would use prior year status to assess growth between the end of the prior year and the end of the first or second quarter of the current year. This growth measure would be used to project additional growth through the end of the year. This model leads to different expected increments across students based on observed differences in rate of growth so far during the year. If any of these models fit satisfactorily, then projections to end-of-year status (and growth) are supportable. If not, then uses of TC1 and TC2 results should be largely unrelated to end-of-year achievement indicators.

References

Connolly, T., & Ordonez, L. (2003). Judgment and decision making. In W. C. Borman, D. R. Ilgen & R. J. Klimoski (Eds.), *Handbook of Psychology, Volume 12: Industrial and Organizational Psychology*. New York: John Wiley & Sons.

Drasgow, F. and Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.

Jones, L. V. & Olkin, I. (2004). *The nation's report card: Evolution and perspectives*. Phi Delta Kappa Educational Foundation.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas, and policy*. Thousand Oaks, CA: Corwin Press, Inc.

Luecht, R.M. and Miller, T.R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16*, 279-293.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

National Center for Education Statistics (2009). *NAEP Technical Documentation*. Retrieved from http://nces.ed.gov/nationsreportcard/tdw/analysis/scaling.asp, on July 27, 2011.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrka Monograph* Supplement, No. 17.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 82-94.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluatin of the National Assessment Governing Board's inaugural effort to set achievement levels on the National Assessment of Educational Progress*. Kalamazoo, MI: Western Michigan University.

Stout, W.F., Habing, B., Douglas, J., Kim, H.R., Roussos, L. & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 19,* 331-354.

Wise, L. L. (2011). *Picking up the pieces: Aggregating results from through-course assessments.* Princeton, NJ: ETS Center for K-12 Assessment.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.