

Guidance for Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems (Extended Version)

Joan L. Herman, Margaret Heritage, and Pete Goldschmidt



Assessment and Accountability
Comprehensive Center

AACC • A WestEd and CRESST partnership

AACC: Assessment and Accountability Comprehensive Center: A WestEd and CRESST partnership.
aacompcenter.org

Copyright © 2011 The Regents of the University of California.

The work reported herein was supported by WestEd, grant number 4956 s05-093, as administered by the U.S. Department of Education. The findings and opinions expressed herein are those of the author(s) and do not necessarily reflect the positions or policies of the AACC, CRESST, WestEd, or the U.S. Department of Education.

A policy brief of this report is available at www.CRESST.org and at www.aacompcenter.org.

To cite from this report, please use the following as your APA reference:

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Developing and selecting assessments of student growth for use in teacher evaluation systems (extended version)*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

The authors wish to thank the following reviewers for their very valuable comments and suggestions on earlier versions that helped us to strengthen this report: Derek Briggs (University of Colorado at Boulder), Michael Gallagher (North Carolina Department of Public Instruction), Neal Gibson, (Arkansas Department of Education), Laura Goe (National Comprehensive Center for Teacher Quality), Lynn Holdheide (National Comprehensive Center for Teacher Quality), and Scott Marion (Center for Assessment). The authors also thank Tamara Lau and Karisa Peer (CRESST) for design and editorial support.



National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

Guidance for Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems (Extended Version)

Joan L. Herman, Margaret Heritage, and Pete Goldschmidt

States and districts across the country are grappling with how to incorporate assessments of student learning into their teacher evaluation systems. Sophisticated statistical models have been proposed to estimate the relative value individual teachers add to their students' assessment performance (hence the term teacher "value-added" measures). The strengths and limitations of these statistical models, as well as the value-added measures they produce, have been widely debated. Little attention, however, has been devoted to the quality of the student assessments that these models use to estimate student growth, which is fundamental to the trustworthiness of any teacher value-added measure.

Assessments that nominally address the subject or grade level that educators teach do not necessarily suffice for the purposes of measuring growth and calculating the value that teachers contribute to that growth. In fact, student growth scores require at least two assessments of student learning - one near the beginning of the school year or the end of the prior year and another at the end of the current school year. Carefully designed and validated assessments are needed to provide trustworthy evidence of teacher quality. Herein lies the purpose of this brief: to provide guidance to states and districts as they develop and/or select and refine assessments of student growth so that the assessments can well serve teacher evaluation purposes.

Applicable across content areas and grade levels, the guidance is grounded in a validity framework that:

1. Establishes the basic argument, which justifies the use of assessments to measure student growth as part of teacher evaluation
2. Lays out essential claims within the argument that need to be justified
3. Suggests sources of evidence for substantiating the claims
4. Uses accumulated evidence to evaluate and improve score validity

The framework is purposively comprehensive in laying out a broad set of claims and potential evidence, which are intended to support long-term plans to validate assessments. However, we recognize that states and districts must respond to current policy mandates, and

operating under both limited resources and tremendous time pressure, cannot be expected to address the entire framework. Nonetheless, by understanding the basic requirements that student assessments need to satisfy, as well as central assessment design features, we believe that the guidance can help states and districts move forward, accumulating important evidence and making improvements in the quality of assessments.

The Basic Argument Justifying Use in Teacher Evaluation

Validity is the overarching concept that defines quality in educational measurement. In essence, validity is the extent to which an assessment measures what it is intended to measure *and* provides sound evidence for specific decision-making purposes. Assessments themselves are neither valid nor invalid. Rather, validation involves evaluating or justifying a specific interpretation(s) or use(s) of the scores.

The process of justifying the use of student growth scores for teacher evaluation takes the form of an evidence-based argument that links student performance on assessments to specific interpretations, conclusions, or decisions that are to be made on the basis of assessment performance. The argument is set out as a series of propositions and attendant claims requiring substantiation with evidence.

Propositions

The general propositions that comprise the argument are:

1. The standards clearly *define* what students are expected to learn.
2. The assessment instruments are *designed* to accurately and fairly address what students are expected to learn.
3. Student assessment scores accurately and fairly *measure* what students have learned.
4. Student assessment scores accurately and fairly *measure* student growth.
5. Students' growth scores (based on the assessments) can be accurately and fairly attributed to the contributions of individual teachers.

Although the first proposition clearly falls outside of the domain of test development and validation, it is an essential requisite for it. Assessment development and/or selection for purposes of teacher evaluation must be guided by publically available and agreed upon learning expectations and not simply by what is easy or convenient to test.

The second general proposition highlights the importance of sound instrument design, development, and review processes in creating trustworthy measures of student growth; whereas, the third and fourth propositions target psychometric and technical qualities of student scores. The final proposition focuses on the technical quality of the teacher value-added scores, which are generated from the individual student growth scores using complex statistical models. While some would regard this final proposition as beyond the province of test validation, we include it as an essential

part of the validity argument and the ultimate link between the test scores to their intended evaluation use.

Figure 1 displays these propositions as a series of if/then statements, which comprise the argument that justifies the use of student assessments to measure student growth for the purpose of evaluating teachers. The sequence of propositions represents the successive issues that states and districts should attend to as they select, develop, and/or refine measures of student growth to evaluate teachers.

Essential Claims and Evidence

With the propositions laid out, the next step in validation involves establishing claims and evidence sources that are important for evaluating each proposition (see Table 1). Like the propositions, claims are of two basic types: design claims and psychometric and other technical quality claims.

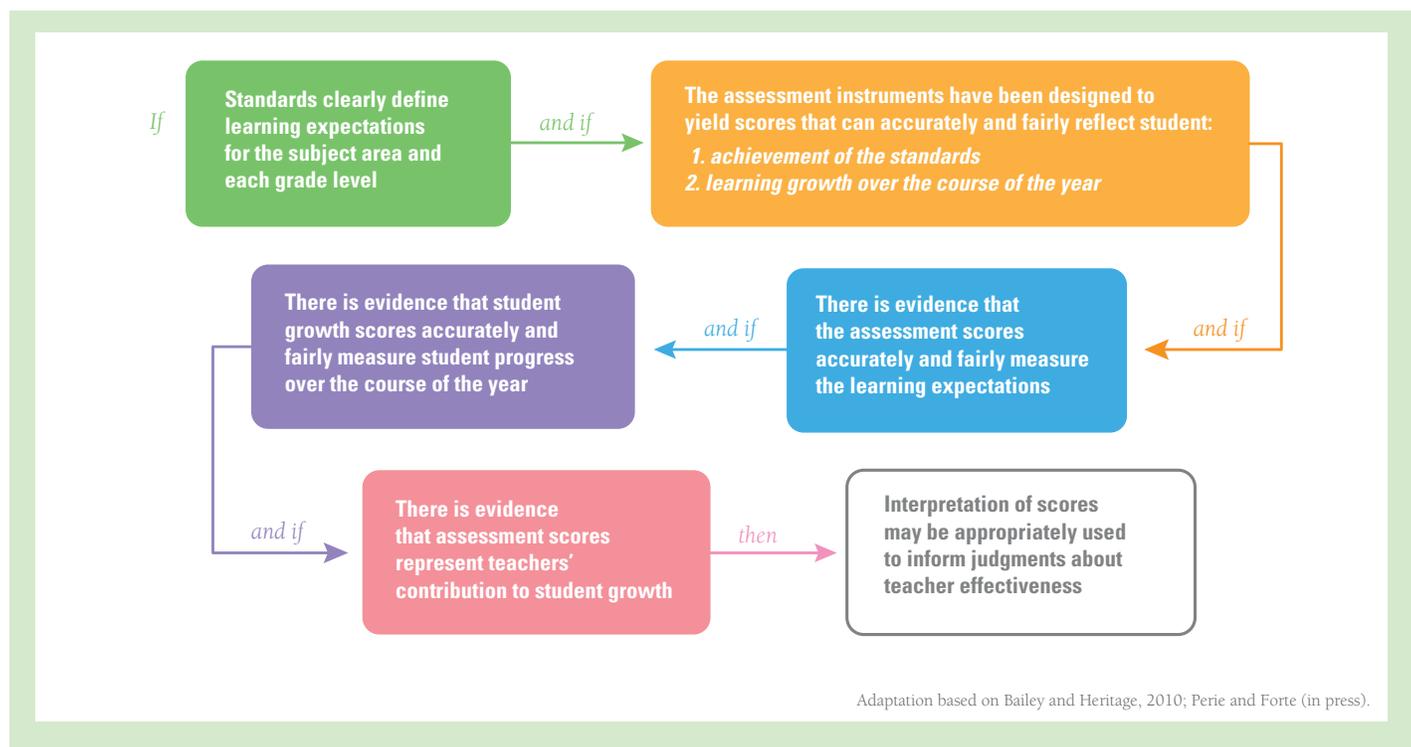


Figure 1. Propositions that justify the use of these measures for evaluating teacher effectiveness.

Design claims. Claims about the attributes and characteristics of the assessment instrument and item design that are likely to yield sound measures. These claims, at least in part, can be examined a priori through evidence produced by rigorous expert review.

Psychometric and other technical quality claims. Claims about the technical quality of the scores and

how well they function as measures of student learning and of teachers' contributions to student progress. The evaluation of these claims draws largely on student data from large-scale field tests or, if necessary, from operational administrations of the assessments and on special research studies that can be coordinated with field-testing and administration.

Proposition 1 - Standards clearly define learning expectations for the subject area and each grade level

Design Claims:

- Learning expectations are clear
- Learning expectations are realistic
- Learning expectations reflect a progression (at minimum for the span of a grade level)

Evidence

- Expert reviews

Proposition 2a - The assessment instruments have been designed to yield scores that can accurately and fairly reflect student achievement of the standards

Design Claims:

- Specifications/blueprint for assessment reflect the breadth and depth of learning expectations
- Assessment items and tasks are consistent with the specifications and comprehensively reflect learning expectations
- Assessment design, administration, and scoring procedures are likely to produce reliable results
- Assessment tasks and items are designed to be accessible and fair for all students

Evidence

- Expert reviews of alignment
- Measurement review of administration and scoring procedures
- Sensitivity reviews

Proposition 2b - The assessment instruments have been designed to yield scores that can accurately and fairly reflect student learning growth over the course of the year

Design Claims:

- Assessments are designed to accurately measure the growth of individual students from the start to the end of the school year
- Cut scores for defining proficiency levels and adequate progress, if relevant, are justifiable
- Assessments are designed to be sensitive to instruction

Evidence

- Expert reviews
- Research studies

Proposition 3 - There is evidence that the assessment scores accurately and fairly measure the learning expectations

Psychometric Claims:

- Psychometric analyses are consistent with/confirm the assessment's learning specifications/blueprint
- Scores are sufficiently precise and reliable
- Scores are fair/unbiased

Evidence

- Psychometric analyses
- Content analysis

Proposition 4 - There is evidence that student growth scores accurately and fairly measure student progress over the course of the year

Psychometric Claims:

- Score scale reflects the full distribution of where students may start and end the year
- Growth scores are sufficiently precise and reliable for all students
- Growth scores are fair/relatively free of bias
- Cut points for adequate student progress are justified

Evidence

- Psychometric modeling and fit statistics
- Sensitivity/bias analyses

Proposition 5 - There is evidence that scores represent individual teachers' contribution to student growth

Psychometric Claims:

- Scores are instructionally sensitive
- Scores representing teacher contribution are sufficiently precise and reliable
- Scores representing teachers contributions are relatively free of bias

Evidence

- Research studies on instructional sensitivity
- Precision and stability metrics
- Advanced statistical tests of modeling alternatives and tenability of assumptions

Based on Herman & Choi, 2010

Table 1. Propositions and Claims Critical to the Validity Evaluation.

Below we elaborate on the claims and evidence for each proposition.



■ **Standards are clear.** First, the standards for each subject, grade, and/or course clearly define important concepts, knowledge, and skills that students are expected to learn as well as the key cognitive demands of each. The performance expectations related to the standard are comprehensively described in terms of what students must know, understand, and be able to do in order to achieve a particular performance level (e.g., proficient versus advanced).

■ **Standards are realistic.** Second, the standards should be *realistic* in terms of the learning that students can be expected to achieve in the course of the year. There also needs to be a manageable number of standards for teachers to address in one grade level, as well as to make sure that deeper learning is not sacrificed for superficial coverage.

■ **Standards reflect progression.** Third, standards must reflect a *progression* of individual student learning, often called vertical alignment. To be vertically aligned, standards must reflect a gross sequence of how related knowledge and skills *progress* from one grade level to the next, clarify how knowledge and skills become wider and deeper in a domain over time, and define what new knowledge and skills need to be learned as students progress across grades. Without such alignment, the knowledge and skills measured in one year may be completely different from those measured in the next. Thus, it would not make sense to compare them directly.

Ideally, the standards will describe a specific progression of how key concepts and skills develop over the course of the year to support the measurement of student progress. A continuous progression of expected learning linked to vertically equated standards provides a critical foundation for instruments that can truly measure what and how much students have learned over the course of a year.

● **Evidence.** Evidence for the design claims for the standards comes from a process of independent review by highly qualified subject matter experts and expert teachers at relevant grade levels. The experts need to consider each standard and/or learning goal for a grade level or course

Even for experts, it is difficult to ascertain what an assessment or item measures simply by looking at it.

and judge 1) how clearly it is described in content and intellectual demand; 2) the extent to which it represents essential learning, (e.g., enduring concepts and principles that are essential for subsequent success in the subject area); and 3) the extent to which the set of grade level standards are feasible for teaching and learning for most students within an academic year. Experts also can rate the extent to which the sequence of standards from one grade level to the next represents a reasonable progression and identify missing and/or non-essential pieces. When experts use structured rating protocols, results can be easily aggregated and used to identify strengths and weaknesses and specific targets for improvement. A number of groups have developed protocols for such ratings, such as Achieve (2008) and AFT (2008).

Grade 5 Informational Text (Key Ideas and Details - R15.2)

Determine two or more main ideas of a text and explain how they are supported by key details; summarize the text

Students develop the skills to:

- identify the main idea of each paragraph (several paragraphs)
- distinguish between important and, less important details in the text related to each main idea
- write a brief statement in students' own words integrating multiple main ideas to identify the significance of the ideas as a whole

Students develop the understanding that:

- a text can contain more than one main idea
- multiple ideas in a text can be integrated
- integrating multiple ideas can show the significance of the ideas as a whole

Grade 6 Informational Text (Key Ideas and Details - R16.2)

Determine the central idea of a text and how it is conveyed through particular details; provide a summary of the text distinct from personal opinions or judgements

Students develop the skills to:

- use text structure and text features to signal central idea of a text
- identify the specific details within the text that convey the central idea
- distinguish fact from opinion in text
- write a summary of a text in students' own words without personal opinion or judgements conveying the central idea of the text

Students develop the understanding that:

- extended text can have a central idea (chapters, entire book)
- particular details of the text convey the central idea
- text structure and text features can assist readers to identify the central idea
- summarizing main ideas requires readers to stand back from what they read and view the text objectively
- personal opinions and judgements are different from objective statements

Grade 7 Informational Text (Key Ideas and Details - R17.2)

Determine two or more central ideas in a text and analyze their development over the course of the text; provide an objective summary of the text

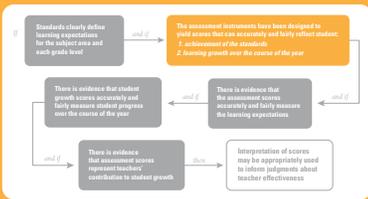
Students develop the skills to:

- objectively summarize the central ideas from a text, preserving the overall meaning
- analyze a summary for evidence of two or more central ideas, supporting details, and objectivity of overall meaning
- trace the development of central ideas in the text, analyzing how the author uses supporting details, text structure, and text features to advance ideas

Students develop the understanding that:

- extended text can have two or more central ideas (chapters, entire book)
- authors use supporting details, text structures, and text features to develop the central ideas over the course of the text
- objective summaries do not contain personal views or opinions of the summary writer

Figure 2: Progression showing intermediate building blocks on the way to meeting Common Core State Standards (ELA Key Ideas).



Proposition 2a

The assessment instruments have been designed to yield scores that can accurately and fairly reflect student achievement of the standards

■ **Specifications/blueprint reflect the breadth and depth of learning expectations.** Specifications provide a blueprint for test development and are a first checkpoint for assuring that an assessment will be well aligned with expected standards. The specifications detail what constructs will be addressed and at what depth of knowledge (see Webb, 2002). For example, will students be expected to define or recall particular concepts, apply and/or explain them, and/or use them to solve complex problems? The specifications also describe the number and types of item formats (e.g., multiple choice, extended response). Additionally, the specifications may include templates to guide item or task development and, where needed, rubrics for scoring student responses.

To measure student growth over the course of a year, specifications should address not only the expected learning progression across a particular grade level, but also include key milestones for prior and subsequent years. Otherwise, a test cannot be entirely mapped to the full range of where some students may be starting at the beginning of the year and to how far some may advance by the end of the year. Without representing the full range, a test may shortchange the actual progress some students have made.

● **Evidence.** Expert subject matter teachers at relevant grade levels and assessment experts need to review the specifications and consider how well they map onto established standards. Reviewers should look for consistency with both the content and depth of knowledge expected by the standards. Authorities on English language learners (ELLs) and students with disabilities (SWD) should also be included in the review process to try to assure that that the design makes the assessments as accessible for as many students as possible. In the case of newly developed assessments for measuring student growth, specifications should be reviewed prior to item development. For existing assessments, the test specifications should be reviewed for information about what is being tested and how well items are aligned with standards.

■ **Assessment items and tasks are consistent with the specifications and comprehensively reflect learning expectations.** While specifications represent the plan for test development, how well the plan is operationalized

is an open question. After items and tasks are initially developed, they typically go through a process of expert and sensitivity reviews and of psychometric testing through pilot and/or field tests. The items and tasks that survive this process (and thus merit being included on an assessment) may not fully reflect the initial plan. Moreover, how the test developer categorizes specific items relative to content and application levels may or may not fully reflect local expectations.

● **Evidence.** Independent subject matter and special population (e.g., ELLs and SWD) experts should conduct an item-by-item review of the operational assessment to examine how well it aligns with the depth and breadth of expected standards. When constructed response or performance assessments are used, it is also important to consider the alignment of scoring rubrics to the standards. In addition, the review should focus on making sure that items are free of elements that would prevent students from some subgroups (e.g., ELL, SWD, gender, cultural) from being able to show their capability. Several well established methodologies exist for conducting such reviews and result in a variety of indices of alignment (e.g., Achieve 2008; Porter, 2002; Webb, 2002).

Because it is difficult to know what an assessment item or task measures by simply looking at it, additional evidence that draws on students' actual responses is desirable. In student think-aloud studies, students are asked to share their thinking process as they complete a task. Their responses are then analyzed to examine what specific concepts, principles, and intellectual processes students actually use to respond. The results of the analysis are then compared to intended standards and/or learning goals. For extended response items, examining student responses can provide a window into the kinds of student thinking and learning that are being elicited by a particular item or task.

■ **Assessment design, administration and, scoring procedures are likely to produce reliable results.** A common rule of thumb is that a minimum of five items is needed to get a reliable score for any narrow test target (e.g., understanding of a particular concept, or ability to apply a specific computational skill). More often, however, a greater number of items is required to get reliable estimates of students' knowledge and skill in a particular

area. For example, an arithmetic assessment containing twenty items may be sufficiently reliable to make inferences about a student's arithmetic ability. Yet, the assessment may contain too few items that specifically target division in order to make reliable inferences about a student's division ability. Generally, increasing the number of items increases the reliability of an assessment (provided that the additional items address the same construct).

Other routine supports for reliability include maintaining test security, standardizing test administration conditions, using a common rubric for scoring, and assuring reliability of scoring. For instance, standardizing the time periods in the school year when students take an assessment, as well as standardizing how long students have to complete the assessments, helps ensure that some students are not unfairly advantaged or disadvantaged by elapsed curriculum time or by available time to complete the assessment.

Consistency in scoring is important for extended response and open-ended items scored by raters. Unless there are clear procedures in place, the scoring of open-ended, extended responses may be subjective, with the result that the scores are too dependent on who does the scoring rather than on the quality of the student response.

The following evidence section enumerates specific elements that should be in place to support scorer agreement and to minimize the error that human scoring can introduce. If raters are not consistent, they introduce additional sources of error.

● **Evidence.** The test and its administration and scoring procedures need to undergo expert review by assessment and measurement specialists. These specialists should examine the assessment directions, booklets, scoring manuals, and other materials to check that sufficiently standardized procedures and conditions are in place to support adequate reliability of scores. Reviewers should check for adequate numbers of items and evidence of reviews for item quality.

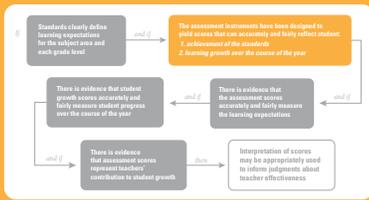
For performance assessment and extended constructed response items, reviewers also should examine the adequacy of scoring and training procedures. For example, clear scoring procedures include specific, well-defined scoring criteria that directly link to the performance expectations of specified standards. Scoring manuals need to include rubrics that operationally define the scales to be used in scoring as well as each score point on the scale. Furthermore, scoring manuals should provide anchor papers or anchor performances that exemplify each score point.

Effective scorer training provides opportunities for scorers to understand, apply, and receive feedback on scoring accuracy. In fact, they should be required to demonstrate their scoring consistency before starting the scoring process. Procedures for checking consistency throughout the scoring process and for re-training scorers who drift from established standards also are important.

■ **Assessment tasks and items are designed to be accessible and fair to all students.** Assessment instruments should permit students to show what they know and should be accessible for all students (including subgroups) for whom the assessment is intended. Test bias arises where there are systematic differences in how students from particular subgroups perform on a test not because of the construct the test is intended to measure but because of construct irrelevant difficulties in assessment tasks and items. In the assessment design phase, care should be taken to avoid construct-irrelevant difficulty in items or tasks that may prevent students from showing their capability. For example, word problems in mathematics that are unnecessarily complex linguistically can present obstacles to ELLs and may prevent them from demonstrating their full capability. Additionally, test items that present stereotypes (of girls, boys, or particular cultures), that portray certain groups as inferior, or that contain images or situations that are more familiar to one group than another could adversely affect the performance of some students. When test bias is present, scores do not have the same meaning for individuals from different subgroups.

To avoid test bias and help ensure that the test will be accessible and fair for all intended subgroups, assessment developers need to clearly understand what constructs are to be measured and to keep in mind the diverse characteristics of students who will take the test. Assessment items and tasks also need careful review to identify and minimize any unintended obstacles (see, for example, Thompson, Johnstone, Anderson, & Miller, 2005; Thurlow, Quenemoen, & Lazarus, 2011).

● **Evidence.** Items should undergo expert review by specialists in English language learners and language development, students with disabilities, and diverse cultural groups. Reviewers should identify and (when necessary) recommend changes or removal of items that contain (a) irrelevant sources of difficulty or unintended obstacles for some students, (b) potentially offensive content, and (c) contexts that may be more familiar to students in some subgroup(s) relative to those in others. As with alignment reviews, standards procedures exist for conducting fairness and bias reviews (e.g., Hambleton & Rodgers, 1995).



Proposition 2b

The assessment instruments have been designed to yield scores that can accurately and fairly reflect student learning growth over the course of the year

■ **Assessments are designed to accurately measure the growth of individual students from the start to the end of the school year.** A student learning growth, in common parlance, denotes the extent to which students gain specific capability from one time point to the next (e.g., Does Marco read better this year than the last? How much did Angela’s math knowledge improve?). Assessments that can be used to measure what and how much individual students learn from one year to the next, as mentioned in design claim 1.0, must be based on vertically aligned standards that describe gross progression of how students’ knowledge, understanding, and skill in a subject matter are expected to develop from one grade to the next. Moreover, to give an accurate portrayal of each student’s progress, instruments should be designed to represent the range of where on the progression students at a particular grade and/or course are likely to start and end over the school year. Of course, the lowest performing students at the beginning of the year may not have attained the expected standards for the prior years and the highest performing students at the end of the year may well have moved beyond the knowledge and skills represented by their grade level’s standards. This range will need to be addressed in instrument design and development. Items and tasks comprising the instruments intended to measure student growth should reflect the full progression of where students are likely to start and end.

● **Evidence.** As with the other claims under this proposition, evidence can be derived from expert review of the items and tasks comprising the assessment(s). Subject matter and teacher experts, along with experts in special populations, can evaluate the extent to which assessment items and tasks reflect vertically aligned standards; in the absence of such standards they can evaluate the extent to which the assessment items present a reasonable progression of development and represent the full range of where different students are likely to start and end the school year. As with other claims requiring expert review, structured protocols will be useful and those previously cited can be augmented to address these additional claims.

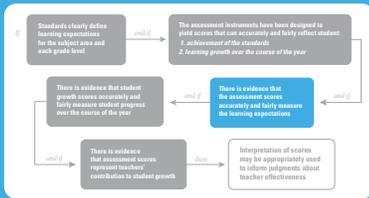
■ **Cut scores for defining proficiency levels and adequate progress, if relevant, are justifiable.** Just as standards need to be vertically aligned, the various competency levels

(e.g., basic, proficient, advanced) also should make sense within and across grade levels. This means that assessment specifications, items, and tasks must not only represent what counts as “proficient” at each grade level, but also reflect the knowledge and skills representing each competency level from one year to the next. For example, what it takes to be classified as proficient in social studies in grade 6 should require more expertise in terms of understanding and depth of analysis than proficiency in grade 5. Similarly, what constitutes basic and advanced levels of competence for a subject and grade level needs to be represented on the assessments and make sense from one year to the next.

● **Evidence.** As with the other claims under this proposition, evidence can be derived from expert review of the items and tasks comprising the assessment(s). Subject matter and teaching experts, along with experts in special populations, can evaluate the knowledge and skills that reflect each competency level and the extent to which these are represented in the assessment.

■ **Assessments are designed to be sensitive to instruction.** Because student growth measures are intended to assess learning growth as a result of instruction, they must, as far as possible, be designed to be instructionally sensitive. Assessments that are likely to be sensitive to instruction are composed of items and tasks that reflect the core goals represented in standards and learning progressions and do not include tangentially related content. Item sensitivity, according to Popham (2007), can be gauged by the extent to which a student’s likelihood of responding correctly would be determined by the student’s (a) socio-economic status or family background, (b) inherited aptitudes (i.e., innate verbal, quantitative and/or spatial abilities), and (c) responsiveness to instruction. The latter criterion is key: if a teacher has provided reasonably effective instruction on the targeted learning goals, is it likely that most of his or her students will respond correctly to the item.

● **Evidence.** Independent review of assessment items and tasks by subject matter and teacher experts can provide evidence of sensitivity of the test design.



Proposition 3

There is evidence that the assessment scores accurately and fairly measure the learning expectations

■ **Psychometric analyses are consistent with the assessment’s specifications.** Test specifications lay out a plan for measuring students’ performance on standards by mapping both the content and depth of knowledge that are to be addressed by the test items. For longer tests, the test specification also may organize standards by strands (e.g., in mathematics, basic computation, applications, problem solving) and provide separate subscale reports for each. Psychometric analyses – the empirical analysis of the item and test scores – can tell you, in part, how well that plan has been realized. These analyses include item statistics such as the difficulty of individual items (what percentage of students got individual or sets of items correct or scored at different levels), indicators of how well the items cohered, and how many separable dimensions or factors are evident in students scores, various indices of precision and reliability as described further below, etc. These statistics can reveal patterns that can be used to evaluate the consistency between a test and its specifications. For example, if test items are written at three levels of depth of knowledge, in general you expect the level 3 items to be more difficult for student than the level 1 and 2 items. If they are not, the items may not be measuring what they were intended to measure. If the test is designed to produce separate subscores (e.g., vocabulary and inferential comprehension), the psychometric analyses – in this case dimensionality analyses – should confirm the existence of these separable dimensions; items intended to measure each should cluster together within and across grade levels.

● **Evidence.** Basic item statistics and dimensionality analyses reveal how a test is operating. A variety of these statistics are generated by standard software programs for test analysis. Evaluating the consistency between these statistics and the test plan may require close cooperation between those with substantive knowledge of what the test is supposed to measure – assessment designers and content experts – and psychometricians, who are knowledge about what patterns are evident in item and test responses and how to look for them. Both perspectives are necessary to get a complete picture, and where necessary to identify where and how the test and/or specific items need to be refined.

■ **Scores are sufficiently precise and reliable.** Precision is an essential characteristic of assessment scores. All test scores are only estimates of what students know and can do, in part because of errors introduced by the measurement process.

Because of these errors, measurement theory assumes that any score a student attains is only an estimate of his or her “true” score on the test - the score the student would attain if there were no measurement error. The concept of “standard error of measurement” provides a gauge of how much error a score contains and is used to calculate the range within which students’ “true” scores are likely to fall. This range – also known as a confidence interval– is important because sometimes the band over which a student’s score likely falls may cross more than one competency level. For example, the confidence interval around a student’s obtained score may include scores that would classify the student as proficient as well as those that would classify the student as basic – making it impossible to convincingly classify that student as being at one level or the other.

Reliability, which is the consistency or stability of students’ observed scores, is typically established through one of three major theoretical approaches: Classical Test Theory (CTT), generalizability theory (G-Theory), and/or Item Response Theory (IRT). In CTT reliability of scores is indicated by the consistency of test scores across different items. G-Theory is applied to tasks, scorers and/or occasions to examine potential sources of error in scores. The basic idea of reliability in CTT and G-Theory is that if an assessment score is to be meaningful, it must measure something stable that transcends the specific time an assessment is administered, the specific test items or test form used, and for constructed response and performance tasks, who does the scoring. Otherwise, it is hard to make the case that test scores reflect a stable capability that merits attention. A variety of reliability statistics examine the relationship between students’ scores across these various assessment conditions. In all cases, the indices range from 0 to 1—the higher the coefficient, the more reliable the test.

In contrast, drawing on concepts of item and test functioning and complex mathematical modeling, IRT offers the important advantage of being able to estimate

the reliability of scores for students at different ability levels or at different score levels. Although not the main mechanism that IRT advances, the measurement information it provides about item functioning can be used to create tests that maximize test reliability of scores around a particular cut point, such as the one used to determine whether students are proficient.

● **Evidence. Precision.** The standard error of measurement (SEM) is the usual statistic used to evaluate measurement error and thus to gauge precision. The SEM, along with a specified level of confidence (typically, 95% or 90%), is used to calculate the confidence interval. While SEM and reliability are inversely related (the higher the reliability, the lower the SEM) in measuring the status of student achievement, the situation can be different in estimating growth. One important advantage of IRT is that, unlike CTT, SEMs depend on where the score falls on the distribution (e.g., SEMs are much larger in the tails of the score distribution than near the mean).

Reliability. Test manuals may include a variety of reliability indices. Classical Test Theory indices include:

- *Internal consistency* is a gauge of how well multiple items and tasks measuring the same construct cohere. There are a variety of indicators of internal consistency, including Cronbach's alpha, split-half, and KR20 coefficients. Common rules of thumb are that these indices should be in the vicinity of .8 and above for group decision and .90 and above for individual decision (see George & Mallery, 2003).
- *Test-retest reliability* is where a test is administered to the same group of students on two occasions, with no intervening instruction, and the relationship between the two scores examined. Test-retest reliability should be in the region of .7.
- *Scorer consistency* indicates the extent to which different scorers give the same or similar scores to student responses for constructed response and performance items. Indicators of scorer consistency include rater agreement, generalizability, Cohen's Kappa.

The *separation index* in IRT corresponds to internal consistency in Classical Test Theory. However, IRT provides additional indices such as marginal reliability, item and test information functions, and model fit statistics that provide additional evidence of how consistently items and tests differentiate students of different ability.

While we have provided some rules of thumb for evaluating reliability statistics, we also recognize that required levels of reliability depend on purpose. Higher levels of reliability are required for high-stakes purposes;

thus, levels should be higher when inferences are made about individual students or teachers than when making inferences about groups.

■ **Scores are fair/unbiased.** Psychometric analyses are used to examine the extent to which items, tasks, and the assessment as a whole function similarly for all students. Where differences exist, items are marked for further scrutiny, revision, and/or elimination.

In the same vein, the indices of reliability and/or precision of scores should be similar for students from different subgroups. Furthermore, they should be similarly related to scores from other measures. For instance, the relationship between scores and subsequent success in the next grade should be similar for English language learners, low socioeconomic status students, and students from different ethnic groups.

● **Evidence.** Differential Item Functioning (DIF) analyses are used to identify items that are functioning differently for different subgroups and thus should be subjected to further review. The presence of DIF is not sufficient to indicate bias. However, when subsequent review finds that the DIF seems related to difficulty that is not part of the construct being tested (e.g., language proficiency), then bias is indicated and the item needs to be revised or eliminated.

In addition to DIF, fairness is supported when there is evidence that scores have the same meaning for individuals from different subgroups. For example, the reliability of scores and SEMs should be computed and compared for all major subgroups, and convergent and divergent relationships should be similar (see below for an explanation of convergent-divergent relationships).

■ **Proficiency Classifications are Consistent and Accurate.** Precision and reliability are important because they impact the consistency and accuracy of performance-level classifications. Classification consistency reflects the extent to which students are similarly classified when they take the same test or parallel forms of it on two different occasions (assuming that no relevant instruction has occurred during the interval between the two administrations). Hence, if students are classified as proficient based on scores from the first administration, they should be similarly classified based on scores from the second. Similarly, the proportion of students who fall into each proficiency level should be similar across the two occasions.

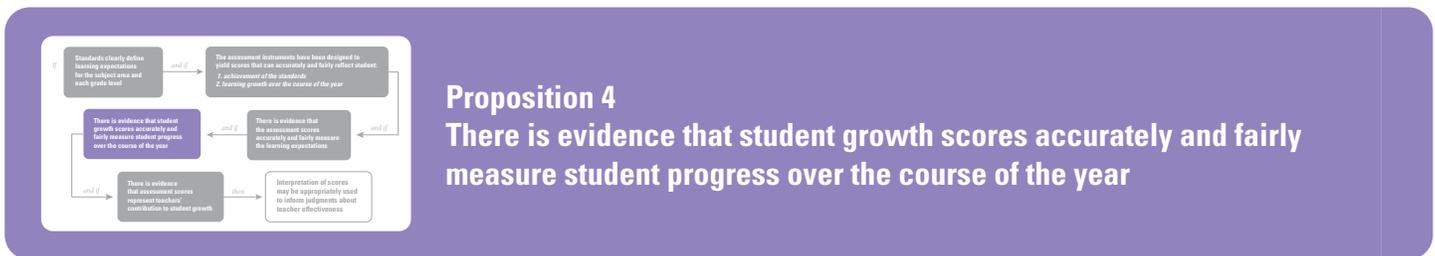
Accuracy in proficiency classification refers to the extent to which a student classified using one assessment would be classified similarly if she could take all possible forms of the assessment. Accuracy determinations draw on Standard Errors of Measurement (SEMs).

● **Evidence.** At the simplest level, evidence of classification consistency and accuracy can be derived from test-retest data to examine the proportion of students who would be similarly classified based on two or more test administrations and/or the probability that students' true scores lie in the level in which they are classified. More generally, determination of classification decisions is based on empirical properties of scores (including the SEM) and predetermined cut points. Several specific strategies exist to conduct the analyses, see for example, Livingstone & Lewis (1995).

■ **Scores are related to other measures of the same construct and less related to measures of different constructs.** If scores from a new assessment, which are intended to measure a given construct, are highly correlated with those from an existing, respected measure of the same construct, then this relationship provides evidence that the new test actually measures what it is intended to measure. Similarly, if assessment scores are supposed to predict who will succeed in college, then the relationship between scores and success in college provides evidence of convergent validity. Such evidence

is strengthened if scores on an assessment not only are highly related to those things it should be related to, but also show less relationship to constructs they are not intended to measure or predict. This general approach is called “convergent-divergent” evidence of validity: it is desirable for scores to converge with other measures of the same knowledge and skills or to future events to which the assessment performance should relate and to diverge from results that tap different domains.

● **Evidence.** To better support score accuracy and fairness, evidence of convergent-divergent validity should be examined. Correlation coefficients should be higher for results from two tests each measuring related constructs or a related future event than with those from tests intended to measure different constructs. For example—if students are administered a new measure of reading comprehension, an existing, respected measure of reading comprehension, and a measure of spelling, one would expect to find higher correlations between scores from the two measures of comprehension than between those from the new measure and the spelling measure. If this is not the case, then it is questionable whether the new measure is accurately measuring reading comprehension.



■ **Score scale reflects the full distribution of where students may start and end the year.** Typically, student progress is measured by comparing students' test performance over subsequent testing occasions (e.g., from the beginning of the year to the end of the year, from grade 4 to grade 5, from grade 5 to grade 6, etc.). However, the direct comparison only makes sense if the scale against which student performance is measured remains constant and has the same meaning across time. Technically, this calls for an equal interval scale, where the meaning of the intervals remains constant across the range of possible scores. For instance, if a student scores 100 in grade 4 and 120 in grade 5, we can infer that the student gained 20 points worth of knowledge and skill only if the two tests are based on the same achievement continuum; if the tests measure similar capabilities from one year to the next; and if the scores have the same meaning from one year to the next (i.e., a fourth grade student taking the fourth grade test and scoring 100 has virtually the same measured capability as a fifth grade student taking the fifth grade test, even though the specific items comprising the test may be different).

Consider the alternative in perhaps a more familiar context: You are on a diet. When you weigh yourself at time 1, you use a scale that measures you in stags. You weigh 50 stags. At time 2, you weigh yourself on a scale that uses drats. You weigh 45 drats. You cannot say you lost 5, because the two measures are on different scales, using different measures. This is why educational test developers rarely report or recommend the use of raw scores. Instead, they convert raw scores into scale scores that reflect a consistent measurement scale from one test form to another or from one year to the next. Think of a measurement scale as a number line that runs from the lowest possible knowledge and skill to the highest level of development; scale scores consistently place students on that continuum.

Ideally, the measurement of individual student growth requires a vertically equated scale score on an interval scale that is comparable across grades—a vertical “yardstick” that runs from the lowest knowledge and skill at the lowest grade level to the highest development

at the highest grade level. It is only in this instance that a change in a scale score from year to year (or from pretest to posttest) theoretically represents an absolute measure of academic progress (irrespective of grade). However, vertical equating is difficult to fully accomplish in practice because it requires that the same constructs be measured from one year to the next. Hence, vertical equating narrows what can be assessed. Moreover, the equating methods that are used usually render scales less comparable as the grade span increases (e.g., from grade 4 to grade 8).

● **Evidence.** Evidence of a good scale begins with documentation that the assessments are aligned with vertically aligned learning standards. Because vertical scaling procedures are complex, as are the indices for

...cut scores that define what constitutes exceptional, adequate, poor, or any other level of progress need to be meaningful and well reasoned.

evaluating scale quality, measurement experts should be asked to review available evidence. In general, such reviews involve examinations of the uni-dimensionality of the scale, grade-to-grade growth patterns and variability, and model fit. Developing sufficient evidence to substantiate an interval scale (as opposed to an ordinal scale) often requires that several psychometricians weigh in. There may be disagreement—even among experts.

■ **Growth scores are sufficiently precise, reliable, and fair for all students.** Beyond being based on a sound, vertical “yardstick,” the precision and reliability of growth scores merit attention. While precision in growth scores is conceptually similar to that of scores from an individual measure, the reliability of growth estimates refers to the ability to detect true differences among teachers based on student growth. If all students’ scores show similar growth, reliability will be very low; thus, the scores cannot differentiate among teachers.

Issues of precision and reliability need to be considered for growth scores for the total population of students and considered separately for students from major subgroups. As with the reliability of scores from a single measure, it is possible that growth scores may be substantially more

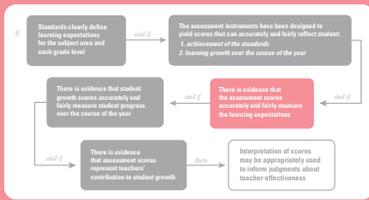
or less reliable and/or precise for some students than for others, and it would be important to take this difference into account in any interpretation. Moreover, differences would have implications for appropriately using growth scores in the evaluation of teachers serving predominantly different subpopulations of students.

● **Evidence.** The reliability and precision of growth scores are evaluated through growth models. While a detailed discussion is beyond the scope of this paper, actual and estimated growth needs to be plotted and compared. The analyses examine the extent to which different growth models actually capture student growth trajectories over time and consider general issues of model fit for the total population of students and for significant subgroups. Models should be checked to see how sensitive or robust they are to differences in specification or to decision rules that determine which data are included (e.g., one year versus two years, or minimum *n* sizes for classrooms or schools).

■ **Cut-points for adequate student progress are justified.** Growth scores are sometimes used to classify students as having made exceptional, adequate, or poor progress. These classifications may be used in calculating the extent of teacher effectiveness with students. As with establishing cut scores for student proficiency levels, cut scores that define what constitutes exceptional, adequate, poor, or any other level of progress need to be meaningful and well reasoned. Those who establish the cut scores should consider as criteria:

- Substantive meaning of the growth scores (i.e., How much and what types of learning does a particular growth score represent?)
- Precision and reliability of the scores relative to the cut points
- Likelihood of achieving various levels of growth (i.e., Based on empirical data, what proportions of students typically achieve each level?)

● **Evidence.** Content and measurement experts should review evidence related to the criteria cited just above to examine whether the different levels of growth reflect substantively meaningful differences, are sufficiently precise and reliable to accurately classify teachers, and reflect likely distributions of performance. Additionally, evidence of actual performance changes should be viewed in conjunction with pre-determined performance levels to determine the extent to which actual growth estimates corroborates the expected movement within and between levels.



Proposition 5

There is evidence that scores represent individual teachers' contributions to student growth

■ **Scores are instructionally sensitive.** Changes in test scores need to reflect the quality of teaching and school programs more than students' innate ability or home background. While we have suggested design characteristics that may help develop tests that will be instructionally sensitive, their actual sensitivity needs to be verified by research evidence. Needed is evidence that well-instructed students outperform students who are not well instructed or those who have had little or no exposure to the assessment targets.

● **Evidence.** A variety of studies can provide evidence of instructional sensitivity by focusing on the relationship between the quality of students' opportunity to learn the knowledge and skills addressed by an assessment and students performance on the test. Students who have more opportunity to learn as well as better quality learning should outperform those who do not. The ideal study would randomly assign students to expert teachers who deliver high quality instruction on tested constructs and to no-instruction groups, and then administer tests to both groups. Practical and statistically significant differences favoring the expert group would provide strong evidence of instructional sensitivity. Alternatively, a state or district could administer surveys of students' opportunity to learn the knowledge and skills addressed by the test in concert with test administration. Controlling for initial differences, students with more opportunity to learn should score higher than those with less.

■ **Scores representing teacher contributions are sufficiently precise, reliable, and free of bias.** Using students' achievement growth to evaluate teachers implies that teachers have directly caused the growth. Yet direct causal inferences ideally require true-experimental studies where students are randomly assigned to teachers. Unfortunately, such conditions do not exist in practice. Instead, complex statistical models are used to estimate and attribute student growth to teachers. The specific model used should be appropriate for the assessments, scales, time frame, and inferences desired. Value added models (VAMs), a subset of growth models, use student achievement growth in various ways to estimate teacher effects.

Teacher effect estimates should be precise enough to distinguish meaningful differences in teacher effects on

student learning. Estimates of teacher effect precision provide such information (e.g., Is there a meaningful difference in student growth between teachers who are classified as highly effective from average teachers or between average teachers and those classified as low performing or ineffective?). Note that reliability must be sufficient to adequately detect true differences in teacher performance. If all students grow at the same rate, then it is not possible to distinguish teachers based on growth (and the corresponding reliability will be very low).

An additional and important form of reliability is stability, which in this case reflects the stability of teacher estimates over time. Evidence should be collected to examine the proportion of teachers who are ranked similarly from one year to the next.

The reliability and precision of teacher growth estimates relate to an overall concern for fairness in using student growth to evaluate teachers. It also is important to consider factors that may influence the estimates but that are not under teachers' control, nor due to their effectiveness. These include:

- *Selection effects*, which result from differences in the students who are assigned to each teacher. For example, if low ability students tend to show less growth than average students, teachers who are assigned predominantly low ability students could be disadvantaged unfairly.
- *Ceiling effects*, which occur when students score at the high end of the scale and do not have sufficient room to show growth from one time period to the next.
- *Spillover effects*, which occur in middle school and high school (as much as 10%) because students' test performance is actually influenced by multiple teachers (e.g., a mathematics teacher also may influence a student's performance in chemistry; in fact, any teacher's estimated effect may be partially a function of the effectiveness of her colleagues).
- *Other measurement factors*, which are a whole range of issues related to the underlying assessments being used. For example, individual assessments may not be sufficiently instructionally sensitive, scaled properly, nor sufficiently reliable to allow for a teacher to demonstrate gains. Additionally, if the same assessments are not used to evaluate all teachers

at a given grade and subject, some teachers may be inappropriately compared to others based on non-comparable assessment results. Assessments following the design and evidence claims outlined in this guidance are likely to yield more meaningful estimates of teachers' contribution to student learning than those that do not.

● **Evidence.** Evidence of the precision, reliability, and fairness of teacher scores will need to be generated by experts in measurement and statistics. Moreover, additional experts should review the specific models used. The meaningfulness and robustness of estimates should be examined in light of reliability, precision, and stability data, as well as with regard to concern for selection, ceiling, spillover, and other measurement effects.

● **Accumulated Evidence to Evaluate Validity.** In sum, we have presented the propositions and claims that constitute the validity argument justifying the use of student growth scores for teacher evaluation. The argument requires a variety of evidence sources, the engagement of a variety of experts, and a long-term agenda.

While this guidance has differentiated design claims from psychometric claims (and the types of evidence on which each is based), the two sources clearly are reciprocally related. On the one hand, the design claims provide the foundation for the technical qualities referenced in the psychometric claims. On the other hand, the psychometric evidence can identify weaknesses in the assessment design that need further refinement.

In fact, both kinds of evidence are used in tandem to identify and respond to potential challenges in the meaning and comparability of assessment scores. Fairness, for instance, is always a central concern in assessment development and validation. The concept of Universal Design means that an assessment is designed to keep the characteristics of all the students in the intended test population in mind. Universal Design is developed to be accessible to diverse individuals and groups (e.g., English language learners, students with disabilities, culturally diverse students). Once developed, items and tasks also are routinely subjected to sensitivity reviews prior to field-testing or operational use in order to identify and remedy potential bias problems. Even so, psychometric analyses may well uncover some items that appear problematic or function differently for students from different subgroups and thus need additional review and possible revision.

Whether based on a full argument or only on selected claims for which data are available, validity evidence is likely to show areas of strength and weakness. Furthermore, validity evidence could suggest areas where assessments may be strengthened to better serve

teacher evaluation purposes. The appraisal may also raise issues where additional evidence is needed. Validation, in short, is a matter of degree and an iterative process that serves both to build the case for the use of the assessment and support improvements in assessment design, interpretation, analysis, and use.

Conclusion

This brief has identified an extensive set of propositions, claims, and evidence sources that are important to the validity argument and which justify the use of student growth assessments as part of teacher evaluation. The set is aspirational; hence, we expect the validity argument to unfold over time.

Under strong policy mandates, many states and districts have had to adopt aggressive timelines for implementing teacher-evaluation systems that incorporate student growth as a component for all grades and subjects. This rapid press for implementation means that it is unlikely that the student growth measures used in the early stages of an evaluation system's implementation will meet all (or even many) of the criteria laid out in this brief. Nonetheless, we hope that this guidance will help states and districts reflect on the major areas of concern and initiate a long-term, systematic process to develop relevant evidence, evaluate strengths and weaknesses, and improve the assessments they adopt.

States and districts can utilize the initial propositions and attendant claims to guide their assessment selection and/or development processes. Moreover, they can use the full set to establish a continuing validation agenda. As the sequence of propositions indicates, states and districts should start by establishing clarity about learning expectations and ensuring, as best they can, that selected or developed assessments are well aligned with those expectations and do not contain fatal design flaws. If necessary, evidence for evaluating subsequent propositions can be collected and analyzed in concert with the assessments' first and subsequent operational administrations.

For instance, states and districts can use the design claims and evidence from expert reviews—along with any available technical data related to the psychometric claims—to systematically evaluate and select the best available options from existing assessments. They can use this evaluation, along with the strengths and weaknesses it identifies, to refine the assessment. Over time, additional evidence can be collected to evaluate a fuller set of claims and used(if needed) to further improve the measures. Just as educators are expected to use evidence of student learning to improve their practice, so too should we

expect states and districts to use evidence of validity to improve their use of student growth measures for teacher evaluation.

Finally, we underscore that no assessment, including student growth assessment, is free of error and all are imperfect. *The Standards for Educational and Psychological Testing* (1999) highlights that no important decision should be based on the results of a single assessment because one evaluation cannot adequately capture the multi-faceted domain of teacher effectiveness. Therefore, multiple measures are essential. Assessments of student growth must be as good as possible; yet, we must keep in mind that they are only one part of a sound teacher evaluation system.

References

- Achieve. (2008). *Overview of Achieve Quality Review*. Retrieved from etc.usf.edu/.../Achieve%20QR1%20Criteria%20with%20Examples_Florida_10-28-08.doc
- Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- American Federation of Teachers. (2008). *Sizing up state standards*. Retrieved from <http://www.aft.org/pdfs/teachers/sizingupstandards0308.pdf>
- Bailey, A., & Heritage, M. (2010). *Washington state English language proficiency assessment foundations document*. Evaluating the Validity of English Language Proficiency Assessments Project (EVEA; CFDA 84.368).
- Blank, R. K. (2010, June). *State growth models for school accountability: Progress on developing and reporting measures of student growth*. Washington, DC: Council of Chief State School Officers.
- Braun, H., Chudowsky, N., & Koenig, J. (Eds.) (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: The National Academies Press.
- Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Thomson Wadsworth.
- Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, 2, 217-233.
- Freedle, R. O. (2010). On replicating ethnic test bias effects: The Santelices and Wilson Study. *Harvard Educational Review*, 80, 394-404.
- George, G., & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference*, 11.0 update. Boston, MA: Allyn & Bacon.
- Hambleton, R., & Rodgers, J. (1995). Item bias review. *Practical Assessment, Research & Evaluation*, 4(6). Retrieved from <http://PAREonline.net/getvn.asp?v=4&n=6> .
- Herman, J. L., & Choi, K. (2010). *Validation plans for Gates-funded assessments English-language arts and mathematics*. Los Angeles, CA: CRESST.
- Jaeger, R. M. (1993). *Statistics: A spectator sport* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179-197.
- Perie, M., & Forte, E. (in press). Developing a validity argument for assessments of students in the margins. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques*. Charlotte, NC: Information Age Publishing.
- Popham, W. J. (2007). *Conducting instructional sensitivity reviews of educational accountability tests*. IOX. Retrieved from <http://www.ioxassessment.com/download/ConductInstructSensReviewsofEducAccountTests-Jim%20version.pdf>
- Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31(7), 3-14.

- Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments*. (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M. L., Quenemoen, R. F., & Lazarus, S. S. (2011). *Meeting the Needs of special education students: Recommendations for the Race-to-the-Top consortia and states*. Washington, DC: Arabella Advisors.
- Webb, N. L. (2002, December). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.