Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems

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

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.

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

The authors 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

Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems[1]

Joan L. Herman, Margaret Heritage, and Pete Goldschmidt

Across the country, states and districts 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; yet, little attention 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 purpose 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 in order 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 intended to support long-term plans to validate assessments. However, we recognize that states and districts must respond to current policy mandates. Thus, operating under both limited resources and tremendous time pressure, they cannot be expected to address the entire framework. Nevertheless, by understanding the basic requirements the student assessments need to satisfy, and the design features that are central, we believe that our 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 in and of 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.

[1]This brief is a shortened version of *Guidance for Developing and Selecting Student Growth Assessments for Use in Teacher Evaluation.* For those who wish to have more details about the contents herein, please refer to the extended *Guidance*.

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 justifying that student assessments can be used 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.

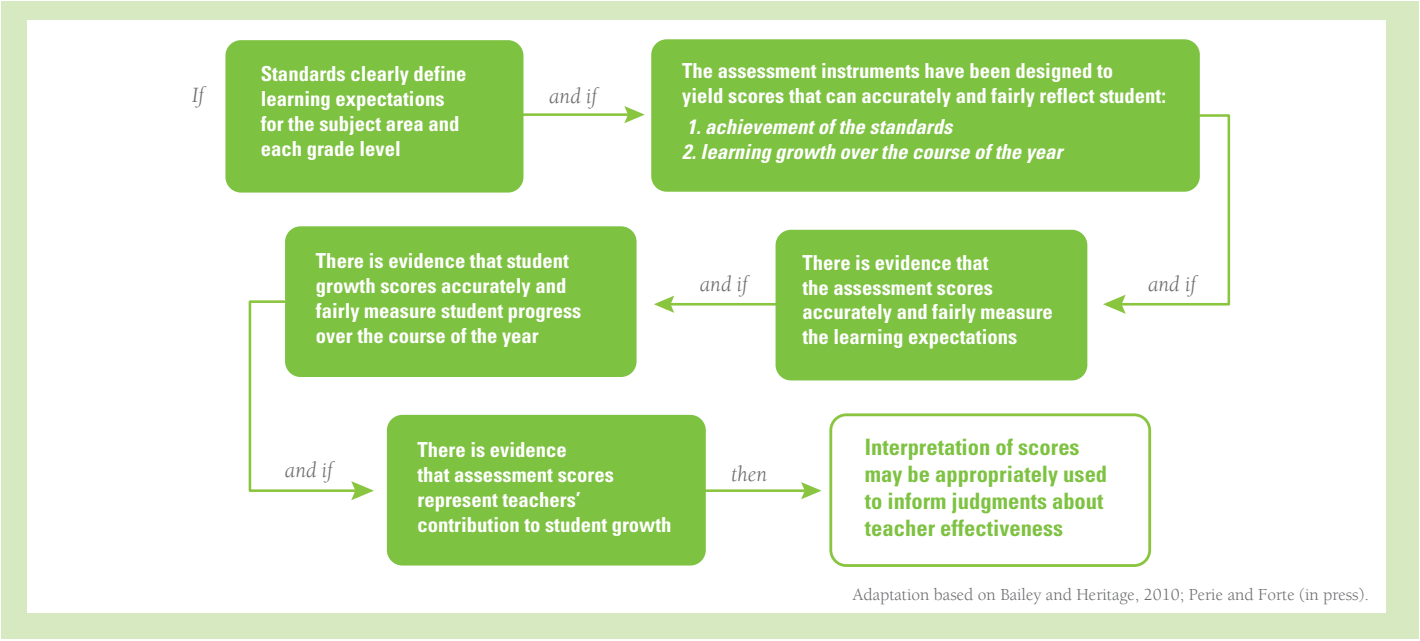

Figure 1. Propositions that justify the use of these measures for evaluating teacher effectiveness.

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: 1) design claims and 2) psychometric and other technical quality claims.

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.

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

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 tenabiliity of assumptions

Based on Herman & Choi, 2010

Expert review. Note in Table 1 that expert review is called for in evaluating claims for Propositions 1 and 2. Highly qualified individuals should comprise review panels—including experts in subject matter, instruction and learning, English learners, students with disabilities, culturally diverse students, measurement and assessment, as well as expert teachers. Their reviews do not all have to be conducted serially; instead, expert panels can convene to conduct reviews simultaneously for many of the design claims for each of the propositions.

The expert panel should engage in structured ratings of the claims, such as those devised by Norman Webb and Andrew Porter. The ratings can be analyzed to provide empirical indices of how well the specifications and actual assessments align with target standards, the frequency of potential bias, and sensitivity or reliability problems. Ratings can also be utilized to summarize what is good, bad, and missing in needed rubrics, administration, training, and scoring procedures. Furthermore, it is often useful to examine the extent of expert agreement: high agreement increases confidence in findings; whereas, low agreement may be cause for concern. The expert reviews provide important feedback that can either be used immediately to strengthen identified weaknesses, or if time is limited, be used in future years to improve assessment quality.

Be aware, however, that expert review has its limits. Even for experts, it is difficult to ascertain what an assessment or item measures simply by looking at it. For performance assessments or expensive assessments in particular—time permitting—it is worthwhile to do small scale think-aloud or cognitive lab studies. These studies ask students to think aloud as they respond to select items or tasks. Student responses are then analyzed to determine whether the tasks actually elicit the content and cognitive demands that were intended, and/or whether the tasks include unintended obstacles preventing some students from showing their knowledge and skills.

Psychometric evidence. As attention moves from design claims to psychometric claims, the demands for specialized measurement and statistical knowledge progressively increase.

The sequence of propositions suggests that the psychometric claims first focus on the individual assessments, which will be used to comprise student growth scores (e.g., assessments given at the beginning and at the end of the academic year). Next, the focus moves to the growth scores and to the teacher value-added measures that are derived from the student growth scores.

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

Problems at an early stage portend larger ones subsequently. Required psychometric and statistical models become increasingly complex as one moves through the continuum from evaluating individual assessments to evaluating the accuracy and fairness of teacher value-added scores. It is likely that states and districts will need to consult measurement and statistical experts to conduct the analyses and review the results (e.g., analysis and review of the specific models used, and the meaningfulness and robustness of estimates with regard to reliability, precision, and stability data).

Reciprocal relationships. Although we have differentiated design and psychometric claims (and the evidence on which each is based), it is important to note the reciprocal relationships between them. On the one hand, the design claims provide the foundation for the technical qualities referred to in the psychometric claims. On the other hand, the evidence related to the psychometric and technical quality claims can identify assessment weaknesses that need further refinement or may raise issues for additional study.

At the same time, the two kinds of evidence are frequently used in concert 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. Applying Universal Design principles during the design phase means that assessment development takes the characteristics of all students for whom the assessment is intended to take into account (e.g., English learners, students with disabilities, culturally diverse students) and helps assure that items and tasks will be accessible to as many students as possible. Items and tasks also are routinely subjected to sensitivity reviews prior to field-testing or operation use. Even so,

psychometric analyses may well uncover some items that appear problematic or function differently for students from different subgroups. These items will need to be re-examined by relevant experts to determine whether a bias problem exists and, if so, whether to eliminate or remedy it.

Accumulated Evidence to Evaluate Validity

Validity is a matter of degree (based on the extent to which an evidence-based argument justifies the use of an assessment for a specific purpose). A complete validity argument, supporting the interpretation and use of growth assessments to evaluate teacher effectiveness, would appraise all of the claims and diverse evidence sources listed in Table 1.

Whether based on a full argument or only on selected claims for which data are available, the appraisal is likely to show areas of strength and weakness and 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 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. As we indicated earlier, the set is aspirational; hence, we expect the validity argument to unfold over time.

Under strong policy mandates, many states and districts 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 aid states and districts to reflect on the major areas of concern as well as 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 on 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, especially 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 utilize 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

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).

Herman, J. L., & Choi, K. (2010). *Validation plans for Gates-funded assessments English-language arts and mathematics.* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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.

Porter, A.C. (2002). Measuring the content of instruction: Uses in research and practice, *Educational Researcher*, *31* (7), 3–14.

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

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education.* (Research Monograph No. 6). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.

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.