National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

GUIDANCE FOR DEVELOPING AND SELECTING STUDENT GROWTH MEASURES FOR USE IN TEACHER EVALUATION

JOAN L. HERMAN, MARGARET HERITAGE AND PETE GOLDSCHMIDT

Introduction

Currently, teacher evaluation is a prominent topic among policy makers across the nation. One aspect of teacher evaluation that is receiving considerable attention is the use of measures of growth in student achievement. The intent of this document is to provide guidance for the development, selection, and/or refinement of student measures that could be appropriate for evaluating teachers' contributions to student learning. Relevant to both tested and non-tested subjects, the guidance focuses on four components that are central to assuring that validity evidence supports the use of assessment results for this intended evaluation purpose:

I. Basic argument justifying the use of student growth measures as part of teacher evaluation.
II. Essential claims of the argument that need to be substantiated
III. Sources of evidence for substantiating the claims
IV. Use considerations
V. Use of accumulated evidence to evaluate validity

The Basic Argument Justifying Use of the Measures

Validity is the overarching concept that defines quality in educational measurement. In essence, validity concerns the extent to which a variety of evidence demonstrates that an assessment measures what it is intended to measure *and* provides sound evidence for specific decision-making purposes.

In modern measurement theory, validation involves first defining an argument that justifies the use of the measures for a specific purpose (Kane, 2004, 2006). The argument is comprised of a series of propositions that link performance on the assessment to specific

interpretations of the meaning of the scores and to specific conclusions or decisions made on the basis of test performance. In the case of student assessments that are used as growth measures for evaluating teacher effectiveness, we see the primary propositions of the argument as:

i. The assessment instruments accurately and fairly measure what students are expected to learn;
ii. The assessments measure accurately and fairly what students have learned over the course of the year;
iii. Student growth based on the assessments can be accurately and fairly attributed to the contributions of individual teachers.

These propositions are laid out in Figure 1 as a series of if/then arguments that articulate the means for reaching the intended end – student assessments that can be used to measure student growth and that can be appropriately used as part of teacher evaluation.

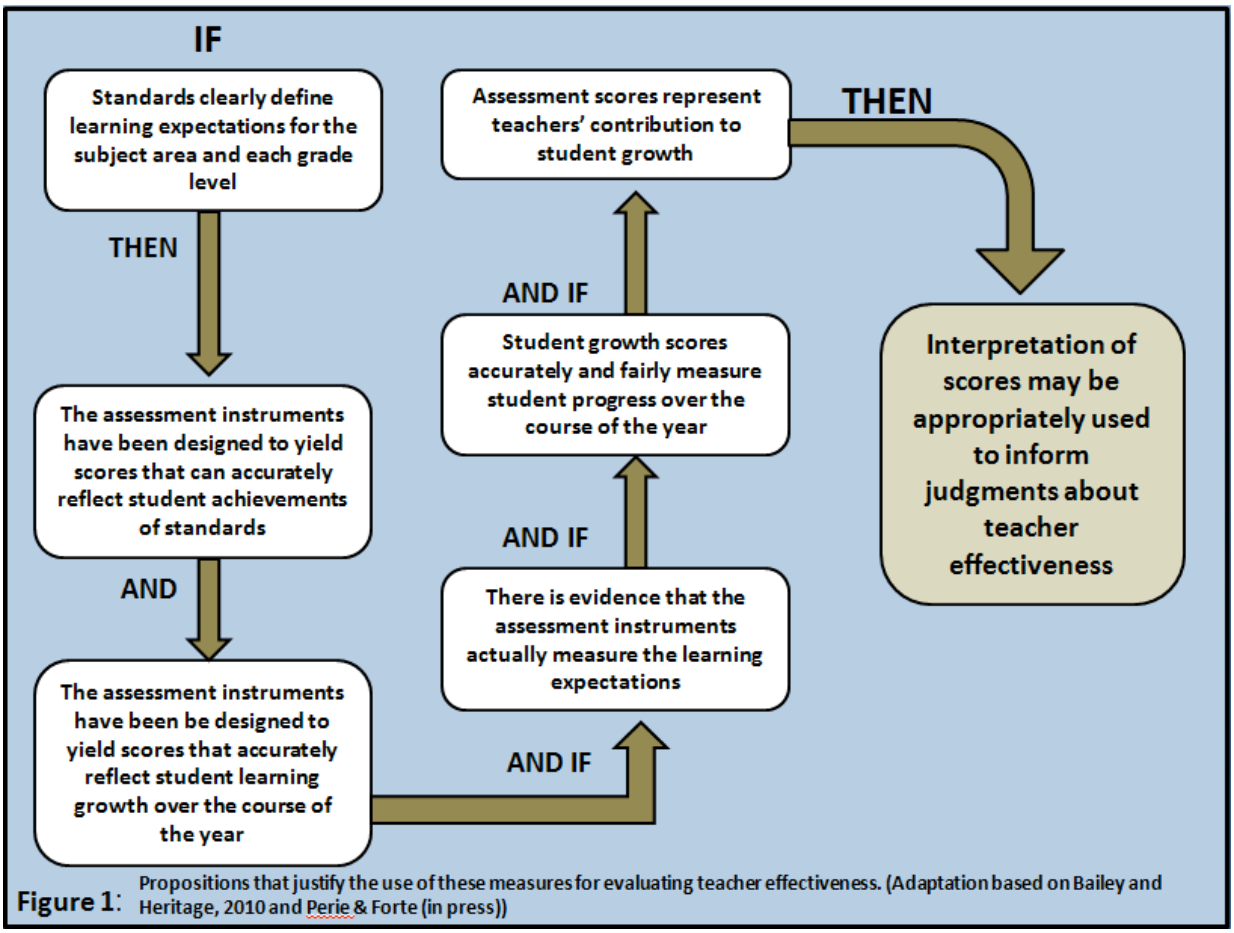

Figure 1: Propositions that justify the use of these measures for evaluating teacher effectiveness. (Adaptation based on Bailey and Heritage, 2010 and Perie & Forte (in press))

The second step involves establishing the claims that support each proposition. These claims constitute fundamental criteria for appraising the extent to which each proposition is supported and needs to be substantiated with specific evidence.

For the propositions in Figure 1 we have identified two primary types of claims:

i) claims about the design characteristics of assessment instruments that may serve the intended evaluation purpose;

ii) claims about the psychometric and other technical qualities that the assessment scores should exhibit to support intended interpretations and use.

The claims for each proposition and potential evidence sources to substantiate them are show in Table 1. Important to note are the reciprocal relationships involved. The design claims provide the foundation for the technical quality of the scores. If evidence shows technical claims are not met, this suggests a return to the design elements so that they can be strengthened.

TABLE 1: PROPOSITIONS AND CLAIMS CRITICAL TO THE VALIDITY EVALUATION

Proposition 1: The standards clearly define learning expectations for the subject area and each grade level.	
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 • content, learning, expert teachers - Research studies validating progressions
Proposition 2: The assessment instruments have been designed to yield scores that can fairly and accurately reflect student achievement of standards.	
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 3a: Assessment scores accurately and fairly reflect the status of students' knowledge and skills relative to learning expectations.	
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 - Logical analysis
Proposition 3b: The assessment instruments have been designed to yield scores that accurately reflect student growth over the course of the year.	
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 4: Student growth scores accurately and fairly measure student progress over the course of the year.	
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: Assessment scores represent teachers' contribution to student growth.	
CLAIMS: - Scores are instructionally sensitive - Scores representing teacher contribution are sufficiently precise and reliable - Scores representing teachers contributions are relatively free of bias	**EVIDENCE:** - Advanced statistical tests (of teacher effects modeling alternatives and collecting empirical evidence assessing the tenability of model assumptions) - Research studies on instructional sensitivity
Based on Herman & Choi, 2010	

It is important that close attention is paid to *all* the design characteristics shown in Table 1 during assessment specification, development and review. Technical evidentiary requirements guide pilot and field-testing. Both claims and evidence provide essential review criteria for examining and/or refining existing tests for potential use in teacher evaluation

Use Considerations

Although not part of the technical evaluation, other issues are important to ensuring the measures well serve their intended purposes as part of teacher evaluation. These issues are represented in Figure 2.

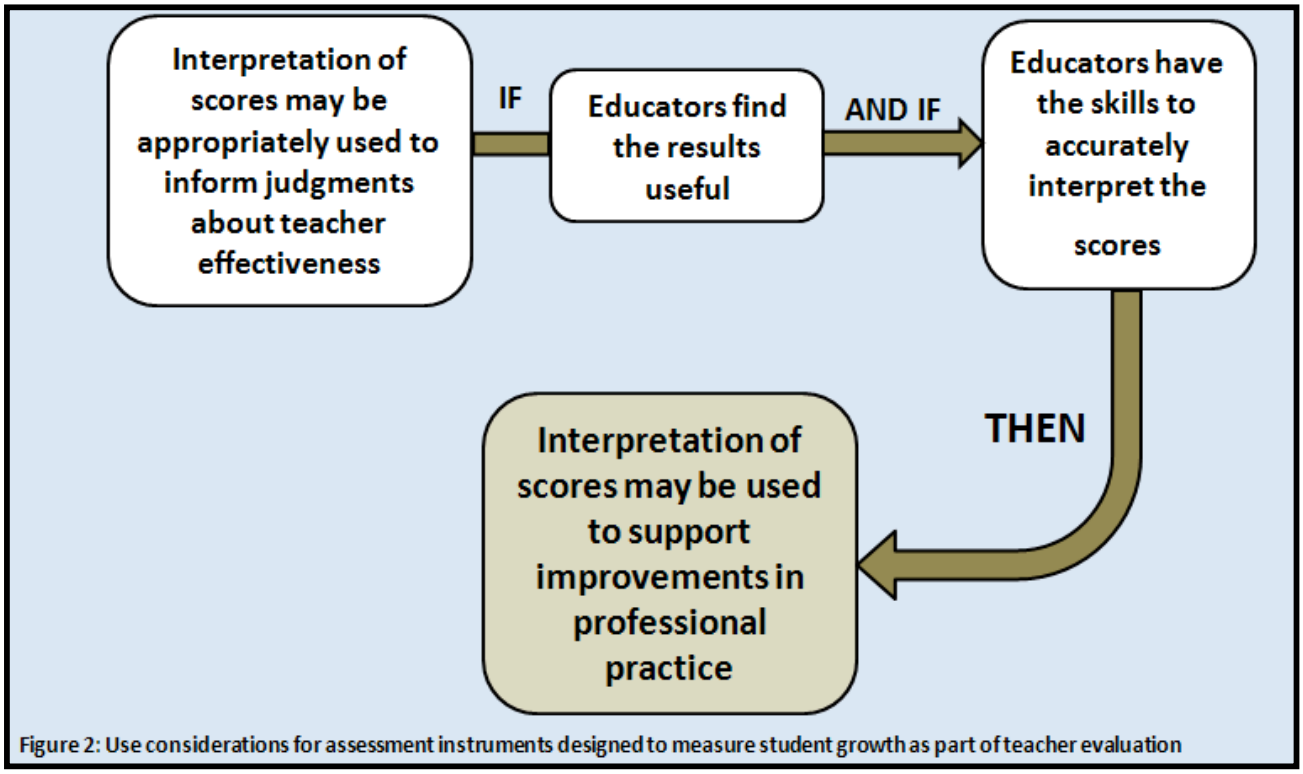

Figure 2: Use considerations for assessment instruments designed to measure student growth as part of teacher evaluation

The measures must be credible and useful to educators. Clearly, if the validity of the measures is not substantiated by evidence, then educators will question their credibility as a component of teacher evaluation. To use the measures effectively to support improvements in professional practice educators must have the necessary skills to interpret the scores and use their interpretations effectively to inform decisions about improving teacher performance. With the necessary interpretive skills, teachers can use the results to reflect on their own practice and engage with peers and administrators to make plans for professional growth. Similarly, administrators will be able to use results to make decisions about teacher performance and professional support if they also have the requisite interpretive skills.

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. Tests themselves are not valid or invalid, rather it is specific interpretations and uses of test scores that are subjected to validation. An assessment may have strong evidence of validity for one purpose but not for another, either because there is limited evidence available or because of what the available evidence reveals. Moreover, it is important to consider each assessment within the broader set that comprises the assessment system and the ability of that system to provide students and teachers equal and fair opportunities to demonstrate performance against consistent consequences.

The validity argument supporting the interpretation and use of growth measures to evaluate teacher effectiveness would appraise the claims and diverse evidence sources outlined in Table 1. Whether based on all such evidence 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 proposed teacher evaluation purposes and to identify areas where additional evidence is needed. An iterative process that builds the case for the use of assessment, validation efforts also can support improvements in the design, interpretation, analysis and use of growth measures for teacher evaluation. Just as we expect educators to use evidence of student learning to improve their practice, so too, should we use evidence of validity to improve our measures.

Finally, no single measure can adequately capture the multi-faceted domain of teacher effectiveness. Regardless of the technical quality of the measures, they should only constitute one part of teacher evaluation. Multiple measures are needed to represent and judge teacher effectiveness.

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: CRESST.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective, 2*, 135-170.

Kane, M. T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.

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