Can Classroom Assessments of Student Growth Be Credibly Used to Evaluate Teachers?

In most parts of America, teachers will soon be experiencing evaluative scrutiny far more intense than anything they’ve ever known. This recent ratcheting up of teacher appraisals can be traced directly to a pair of federal initiatives that have enticed educational officials in nearly every state to shape up—and also toughen up—their state’s teacher-evaluation program. A cornerstone of these federally spawned teacher-evaluation programs is a stipulation that evidence of student growth must be “a significant factor” when judging a given teacher’s quality. A central question flowing from these intensified evaluations is whether teachers can use classroom assessments to collect evidence of their own student growth and, if so, will such evidence be believable?

Federal Carrots and Sticks

Before considering how teachers might evidence regarding their own effectiveness, a bit of context-setting is in order. As noted above, two specific federal initiatives have triggered a massive overhaul of state-level teacher-evaluation programs. In 2009, the Race to the Top program (RTT) offered hefty financial grants to states and, later, to school districts if they were willing to adopt aggressive school-reform strategies—including more demanding teacher-evaluation procedures (“Race to the Top Assessment Program”). Subsequently, in 2011 the ESEA Flexibility program awarded waivers to states so that those states could dodge the increasingly serious sanctions embodied in the most recent version of the Elementary and Secondary Education Act of 1965, that is, the No Child Left Behind Act (NCLB) (“ESEA Flexibility”; “The Elementary and Secondary Education Act of 1965 as Amended Reports to Congress”). As was the case with the RTT, states were more likely to be granted a waiver if they installed teacher-evaluation programs in which student growth—typically determined by students’ test scores—was a significant evaluative consideration.

Because both the RTT’s financial “carrot” and the ESEA Flexibility Program’s “stick-avoidance” called for an emphasis on student-growth evidence when evaluating teachers, it is not surprising that we now find many states’ teacher-evaluation programs requiring evidence about the degree to which a teacher has promoted students’ test-determined student growth. With few exceptions, “student growth” signifies that students will display increased mastery of the skills and bodies of knowledge set forth in a state’s official curricular aims.

For many teachers, student-growth evidence is routinely available in mathematics and reading because of states’ NCLB-required annual accountability tests in grades 3–8 and once in high school. However, for roughly half of the nation’s teachers, no state-administered tests supply students’ annual test scores at certain grades or in particular subjects such as social studies, science, and the arts. Accordingly, to secure evidence of student growth for these teachers, and even for the many elementary teachers whose instructional responsibilities extend beyond math and reading, authorities in many states are counting on teacher-supplied evidence regarding
student growth—typically captured via students’ performances on classroom tests.

Some states have generated a substantial flock of such classroom assessments to be used by their states’ teachers in pursuit of student-growth evidence. Other states are leaving it up to local school districts to work with teachers who must devise and administer their own classroom assessments. Moreover, the evaluative weight given to student-growth evidence collected by classroom assessments varies substantially from state to state. Some states have identified predetermined weightings of classroom-assessment evidence, for instance, 25 percent, while other states leave such weighting decisions to each of its school districts.

A final factor to consider as we view today’s teacher-evaluation landscape is that the need for classroom student-growth evidence differs dramatically among states depending on the degree to which a state’s NCLB accountability tests have been shown to be instructionally sensitive, that is, are accompanied by evidence indicating those tests are actually able to distinguish between well taught and badly taught students. If a state’s accountability tests are unaccompanied by evidence attesting to their instructional sensitivity, then it becomes even more important to rely on classroom-assessment evidence of a teacher’s instructional competence. Put simply, if there’s no evidence indicating that statewide tests are able to accurately evaluate teachers, then more evaluative weight should surely be given to classroom-based evidence of student growth.

The task, then, is for the nation’s teachers to come up with ways of supplying genuinely defensible evidence of their students’ growth, and to do so in a manner that is not only accurate, but is also believable. This, clearly, constitutes a nontrivial challenge.

Two Key Evaluative-Evidence Issues

Clearly, if state laws or regulations require that part of the teacher-evaluation story must be told by using classroom-assessment evidence of student growth, then teachers will want to make certain such evidence is appropriate. But let’s be realistic about today’s demands on most teachers’ time. Faced with financial constraints, many teachers are teaching more students for longer hours and with no increase in recompense. Accordingly, whatever tactics a state’s officials advocate that teachers employ when collecting student-growth evidence, the advocated tactics dare not be so complicated or so time-consuming that most teachers will be reluctant to use those tactics. Rather, what we need is an efficient way for teachers to collect accurate and compelling evidence regarding their students’ growth.

Because few states will be able to afford independent data-collection teams for this purpose, it is almost certain that if classroom-assessment evidence is to play a prominent role in a state’s teacher-evaluation program, individual teachers will be the ones who must provide such evidence. In the collection of classroom student-growth evidence, however, teachers must be attentive to two overridingly important issues: (1) evaluative-inference validity and (2) evidence credibility.

Educational assessment validity refers to the accuracy of score-based inferences about students’ status, for instance, the degree to which students have mastered a particular body of knowledge or a specific cognitive skill. A test measuring students’ ability to interpret tabular data-displays, for example, uses students’ overt responses to the test’s items so teachers can make valid, that is, accurate, inferences about students’ covert data-interpretation skills.

In a teacher-evaluation setting, however, not only do students’ test performances need to yield valid assessment-based inferences about students’ achievements, but those test performances also must allow us to get an accurate fix on a teacher’s quality. In other words, the evaluative inference about a particular teacher’s quality, given real-world limitations, must be as valid as we are able to make it.

The evaluative inference at issue when we appraise a given teacher focuses on a determination of the teacher’s competence. Largely because of federal preferences, most states have already agreed that a teacher’s competence should hinge heavily on the teacher’s ability to promote measurable student growth. Accordingly, we need to increase the likelihood that an evaluative inference based meaningfully
on student-growth evidence will be valid, that is, will accurately reflect a teacher’s competence.

Classroom-assessment evidence bearing on a teacher’s competence can consist not only of students’ performances on traditional paper-and-pencil tests, but can also be collected via a wide variety of assessment ploys such as performance assessments, oral presentations, or student-generated products. Fortunately, in some subject areas such as language arts, students routinely generate products that can be collected sequentially—usually in portfolios—and identified according to when those products were generated. Thus, an English teacher can assemble students’ periodically generated essays so it is apparent whether improvements in those students’ writing skills have occurred. Not all subjects, of course, yield portfolio-collectable student products.

Beyond evaluative-inference validity, a second issue closely allied to those inferences is evidence credibility, that is, the believability of any teacher-collected evidence demonstrating student growth. Candidly, because teachers will often be obliged to collect student-growth evidence to evaluate themselves, skeptics are apt to believe that some teachers might engage in self-serving behaviors intended to make those teachers seem more skilled than they actually are. Accordingly, teachers should attempt to eliminate any reasons for someone to doubt the credibility of classroom-assessment evidence indicating the extent of students’ growth.

Evaluative-Inference Quality

At least four factors can be identified that bear directly on the validity of an evaluative inference about a particular teacher’s quality. Although each of these factors could be treated in substantial detail, a brief description of the four factors will, hopefully, make clear their importance and, therefore, will suggest how they should influence the evaluative weight teacher evaluators should give to any classroom-assessment evidence collected for appraising a given teacher’s quality.

1. Significance of What’s Being Assessed

Educational assessments can be built to measure students’ mastery of what’s significant or what’s trivial. Clearly, we should be inclined to assign greater evaluative weight to classroom-assessment evidence measuring students’ accomplishments that are meaningful rather than trifling. The significance of the knowledge and/or skills being measured by classroom assessments should be evident. In many settings, it should also be made clear that the skills and knowledge being measured by a teacher’s classroom assessments are in alignment with officially approved curricular targets.

2. Quality of the Assessments Being Used

It almost goes without saying that if classroom assessments are being employed to collect data about a teacher’s quality, those assessment instruments should be of high quality. Clearly, evidence garnered by using ill-conceived, hastily constructed tests should be given less evaluative weight than evidence collected by employing thoughtfully devised tests.

Any educational assessment textbook will contain not only a number of guidelines regarding how to construct and improve different kinds of test items, but also descriptions of straightforward ways to collect data regarding the technical quality of a classroom test. The caliber of the assessment instruments being employed in the collection of student-growth evidence is, arguably, among the most important factors to be considered when weighting the worth of such evidence.

3. Two-Occasion Assessment

Student growth, by definition, signifies a change in students’ achievement, that is, a change (hopefully, an increase) in students’ mastery of designated skills and knowledge. To arrive at an accurate determination of a student’s growth, therefore, teachers must collect assessment evidence on two or more occasions. Sometimes, as is true when results of states’ NCLB tests are available, teachers can determine students’ growth by contrasting their students’ performances on this year’s end-of-year tests with those same students’ performances on last year’s end-of-year tests. In most instances, however, to get at “growth” accurately via classroom assessments, teachers will need to rely on some variation of a pre-instruction versus post-instruction design featuring the same test or two essentially equivalent tests.
If a two-test approach is adopted, care must be taken to ensure that the coverage and the difficulty of the tests used are, in fact, sufficiently similar so that apparent “improvements” will not be attributable merely to differences in difficulty between pre-instruction and post-instruction tests. Moreover, attention should be given to the similarity of conditions under which different test administrations take place. For example, the validity of a test-based evaluative inference about a teacher’s quality would evaporate if students had been allowed an entire period to complete a teacher-made posttest, but had been obliged to race through that same test in 20 minutes when, some months earlier, it had served as a pretest.

4. Scoring Inaccuracies

For all educational assessments, but particularly for those containing constructed-response items such as short-answer items or essay responses, scoring accuracy is imperative. If teachers attempt to personally score their own students’ pre-instruction and post-instruction responses, even if those teachers are truly committed to score such responses accurately, it will often be the case that teachers will subconsciously succumb to a wishful-thinking perspective in which they may “see improvements” when, in fact, no such improvements exist. Thus, when a teacher’s evaluation is on the line, efforts to assure that students’ tests have been scored with accuracy should be both made and, ideally, documented.

Involving non-educators, such as students’ parents or members of the community, in objectively scoring students’ responses will typically address this concern nicely. Such nonpartisan scorers can, if necessary, be given guidance about the use of scoring guides, then asked to score students’ responses only after students’ names have been redacted as well as the date on which a test was completed. After this blind-scoring has been completed, previously assigned codes can be used to identify students’ scored responses as pretests or posttests.

Although other factors can have an influence on the validity of the evaluative inference about a teacher’s quality and, therefore, on the weight that should be accorded to specific classroom-assessment evidence regarding a teacher’s quality, these four factors will usually account for most of the reasons we should assign greater or lesser evaluative weight assigned to different incarnations of classroom-assessment evidence.

Evidence Credibility

When supplying evidence regarding one’s own competence—irrespective of a person’s occupation—there is an inherent tendency to present oneself in the best possible light. It’s just human nature.

Accordingly, even the most accurate evidence of a teacher’s quality, if self-submitted, will sometimes be given less evaluative weight by skeptics than such evidence actually deserves. Teachers who collect classroom evidence of their own instructional quality, therefore, should also be attentive to the following four guidelines that, if followed, will enhance the degree of evaluative credibility given to classroom-assessment evidence. Certain of these guidelines will be more difficult to follow than others. Nonetheless, if teachers go to the trouble of collecting evidence regarding their own instructional quality, then it makes sense for those teachers to enhance the credibility of whatever evidence they submit. As you will see, most of the guidelines to be described can be followed either by (1) teacher-supplied documentation or (2) enlisting the aid of independent allies.

Guideline 1: Establish Importance of the Curricular Aims Being Assessed

To make certain that the student-growth evidence being collected deals with curricular aims of appropriate significance, teachers can briefly describe the extent to which the curricular targets being assessed via their classroom tests are in accord with officially approved curricular aims for the grades/subjects involved. In essence, a teacher simply provides a brief written account indicating the alignment between what the teacher’s classroom tests measure and whatever curricular aims the state or local district has prescribed for those students.
Can Classroom Assessments of Student Growth Be Credibly Used to Evaluate Teachers?

If a teacher wished to further support the appropriateness of what’s being measured by the teacher’s classroom tests, an external reviewer might be asked to supply a brief, independent written commentary on the suitability of what’s being measured by the teacher’s classroom assessments. Not only would such an external reviewer (for example, another teacher or a school-site administrator) be asked to comment on the alignment of the teacher’s classroom assessments with officially sanctioned curricular aims, but those reviewers might also be asked to identify the cognitive levels embodied in the tests, that is, the degree to which the tests call for students to engage in higher-order cognitive skills rather than in mere memorization.

Guideline 2: Verify the Quality of the Assessments Employed

Evaluative data collected by superior classroom assessments will almost always be more credible than will data collected by shoddy classroom assessments. Thus, if teachers have engaged in any activities bearing on the technical quality of their classroom tests (such as by analyzing the reliability or, perhaps, the instructional sensitivity of the tests), the results of those activities should be submitted the same time any classroom-assessment evidence is provided to teacher evaluators. Even if no such technical-review activities have been undertaken, actual copies of any tests employed should always accompany the submission of evaluative evidence based on those tests.

Guideline 3: Verify Test-Administration Equivalence

If there are substantial differences in the way that pretests and posttests have been administered to students, then those differences can reduce the perceived persuasiveness of classroom-assessment evidence about a teacher. Differences in the amount of testing time allowed for a pretest and a posttest, for example, can obviously make a difference in students’ performances. And there are other ways in which differences in test-administrations can lead to distortions in the resultant data. For instance, suppose that when students complete a pre-instruction test, they are given little, if any, support from a teacher during the actual test-taking. In contrast, however, suppose that during the post-instruction exam a teacher roams the classroom to supply helpful hints whenever any students appear to be having trouble with certain items. Softened posttest administrations can, clearly, lead to higher scores than warranted.

How, then, might a teacher demonstrate that conditions during two test-administrations are the same? One way is to write up a brief description of the key conditions associated with the two testing occasions, for example, the precise amount of time allowed for each, the amount of teacher-supplied clarifications during the test-taking time, the nature of the teacher’s test-taking directions to students prior to each test, and so on. In short, a teacher simply describes in writing the nature of the two testing occasions in a manner that will permit teacher evaluators to judge whether test-administration equivalence is, indeed, present.

If a teacher believes that the teacher evaluators who will be judging classroom-assessment evidence are apt to be particularly dubious regarding the credibility of such evidence, it is also possible to enlist the assistance of a nonpartisan individual, for instance, a school-site administrator or even a parent, to briefly interview a handful of randomly chosen students about the conditions associated with the two testing conditions. A brief interview guide featuring a few questions such as the following could be employed: “As you recall the time three months ago that you and your classmates completed the pretest version of the posttest you just finished, how similar were the amounts of time you were given to take the two tests?” Clearly, going to the trouble of securing an ally to obtain such evidence independently represents more hassle for a teacher than a personally recounted description regarding the similarity of the two testing conditions. A brief interview guide featuring a few questions such as the following could be employed: “As you recall the time three months ago that you and your classmates completed the pretest version of the posttest you just finished, how similar were the amounts of time you were given to take the two tests?” Clearly, going to the trouble of securing an ally to obtain such evidence independently represents more hassle for a teacher than a personally recounted description regarding the similarity of the two testing conditions. A brief interview guide featuring a few questions such as the following could be employed: “As you recall the time three months ago that you and your classmates completed the pretest version of the posttest you just finished, how similar were the amounts of time you were given to take the two tests?” Clearly, going to the trouble of securing an ally to obtain such evidence independently represents more hassle for a teacher than a personally recounted description regarding the similarity of the two testing conditions.
inference is for teachers to teach directly toward the specific items on a test rather than toward the skills and/or knowledge represented by the test’s items. If, for instance, during the time between a pretest and a posttest, a teacher has supplied gobs of practice time prepping students in how to correctly answer particular items on the posttest, students’ posttest scores might soar, but any inferences about students’ increased mastery of the skills and/or knowledge assessed by the test are likely to be invalid.

A teacher can personally document the degree to which, during the instructional segment involved, any item-focused teaching was present. This can be done simply by the teacher’s describing, perhaps in a paragraph or two, the extent to which instructional activities were aimed at students’ skill-mastery (or knowledge-mastery) rather than at particular posttest items. A flat-out denial of any item-specific instruction might also be registered by a teacher.

In addition, however, if a teacher wishes, it is also possible to enlist the aid of someone who can independently collect interview data from a small sample of randomly chosen students regarding this issue. Students might, soon after completing a post-instruction exam, be asked to respond anonymously to a few questions such as the following: “When you and your classmates were being taught the things measured by the exam you recently completed, were you often taught how to answer the specific items included on the exam—not whether you were taught about the skills or knowledge being measure by the exam, but were you taught to answer correctly the actual items on the exam?”

If a teacher chose to rely on student-supplied reactions to questions dealing with item-focused instruction related to this fourth guideline, those questions could be collected by an independent interviewer along with implementation of other guidelines.

It is always possible that, even if a teacher did not personally choose to collect evidence from students regarding item-focused instruction, such a student-interview procedure might be required by a school’s teacher-evaluation system. If this were done, and if teachers were aware of such a possibility once students had completed a post-instruction test, this possibility would most likely dissuade many teachers from engaging in any item-focused instruction.

**Should Preceded by Can**

This analysis revolves around a fairly simple question: Should teacher-supplied evidence of student growth—evidence garnered by using classroom assessments—be used to evaluate a teacher? Because of the federal government’s strong urging that the evaluation of teachers rely on “multiple measures” of student growth, and one set of those measures can certainly include evidence collected by teachers’ classroom assessments, it is apparent that if reasonably accurate evidence of student growth can be gathered via classroom assessments, then such evidence most definitely should be employed in the evaluation of a given teacher. Indeed, if many of a state’s teachers are to be evaluated according to their students’ growth on state accountability tests for which no evidence of instructional sensitivity exists, then student growth indicated by classroom assessments should be given paramount importance.

As has been pointed out, however, the quality of evidence—and the degree of credibility accorded to such evidence—can range considerably. Not evidence from any old classroom assessment should play a prominent role in a teacher’s evaluation. However, carefully collected classroom-assessment evidence of students’ growth will make a potent contribution to the appraisal of any teacher. To make our nation’s current teacher-evaluation extravaganza have a truly happy ending, it is apparent that teachers, as well as the teacher evaluators who judge them, must become truly knowledgeable about how to collect classroom-assessment evidence of student growth. We need evidence that’s not only accurate but can also be believed.

**Works Cited**


W. James Popham. UCLA emeritus professor, is the author of 2013’s Evaluating America’s Teachers: Mission Possible? by Corwin Press. Email him at wpopham@ucla.edu.