Journal of Advanced Academics
22(5) 750–770
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1932202X11424882
http://joaa.sagepub.com

# Measuring Teacher Effectiveness in Gifted Education: Some Challenges and Suggestions

Megan E. Welsh[1]

## Abstract

States and districts are under increasing pressure to evaluate the effectiveness of their teachers and to ensure that all students receive high-quality instruction. This article describes some of the challenges associated with current effectiveness approaches, including paper-and-pencil tests of pedagogical content knowledge, classroom observation systems, and value-added models. It proposes development of a new teacher evaluation system using a virtual reality environment and describes how innovations in educational measurement and technology can be used to develop an improved teacher effectiveness measure.

## Keywords

teacher effectiveness, assessment, educational technology

A principal and a fifth-grade teacher sit down to discuss the teacher's annual evaluation. The principal presents three classroom observation forms, completed after each visit to the teacher's classroom, and student gains on the annual state assessment. The teacher is familiar with the observation results, having discussed them before. "My observations indicate that you are very adept at differentiating instruction, provide engaging and challenging lessons, and have also built a real sense of community in your classroom," the principal begins.

[1] University of Connecticut, Storrs, USA

**Corresponding Author:**
Megan E. Welsh, Department of Educational Psychology, University of Connecticut, 249 Glenbrook Road
Unit 2064, Storrs, CT 06269-2064, USA
Email: megan.welsh@uconn.edu

> However, I did not see you use any instructional technology at all and, because you were working with small groups of students, it was both difficult to hear and to rate your skills in giving feedback. In addition, your students only gained 5 points on the state test, well below the 15 points gained on average throughout the school. Therefore, while you will receive merit pay this year, I have decided to give you only partial merit instead of the full amount.

The teacher, surprised and somewhat offended by his supervisor's decision, responds,

> That is quite disheartening since my students and I use technology daily, but have not incorporated it into literature circle and have only partially incorporated it into writing workshop, the activities you observed. As you know, both my students and I have done quite well. Most of my students scored above the 90th percentile on the state test, and I scored at the 95th percentile on an exam administered by the National Board of Professional Teaching Standards, which included feedback among the skills it assessed. And, by the way, what level of merit pay did kindergarten teachers with similar observations receive since they can't be evaluated with state test scores?

As the scenario above illustrates, teacher effectiveness is very difficult to measure; evaluation results can be affected both by the measures used and by the way that scores are analyzed. In addition, multiple definitions of effective teaching abound, with some focused on student achievement, some on the classroom experiences teachers create, and others on students' social and emotional development (Brophy & Good, 1986; R. J. Campbell, Kyriakides, Mujis, & Robinson, 2004; Goe, Bell, & Little, 2008).

Discussions like the one above are likely to increase as states and districts move to evaluate teacher's effectiveness using multiple measures. Recent federal education policies award funds for elementary and secondary education on a competitive basis, with teacher effectiveness and the equitable distribution of highly qualified teachers among the key issues that must be addressed in a viable application (American Recovery and Reinvestment Act, 2009). Several states, including Arkansas, Florida, Louisiana, Maryland, Oklahoma, Tennessee, and Virginia, have enacted teacher evaluation systems that include student achievement among the key sources of evidence (Dixon, 2011). As teacher effectiveness becomes central to educational accountability—in terms of merit pay and in response to calls to remove or reform ineffective educators—it is imperative that we find ways to accurately evaluate teachers. Ideally, these evaluations should reflect a teacher's skill with the students he or she serves and result in comparable ratings across districts, schools, and classrooms.

All three current teacher evaluation methodologies—classroom observations, examining student achievement, and teacher tests—have many strengths, and the combination of these approaches is particularly strong. However, these evaluation methodologies may not yield scores that are comparable across classroom settings, and they could lead to sanctions against the wrong set of educators. To avoid this, a new set of teacher effectiveness measures are needed: measures that gauge teacher performance given a uniform set of students and curriculum that is similar to the teacher's current placement. These measures would not replace current approaches but rather supplement them by providing effectiveness scores that control for variability across students, content, and grade levels.

Such an assessment would use technology to create a video game–like simulation to ensure that each teacher experiences not only the same set of students and curriculum but also the exact same student behaviors during their evaluation. As the performance assessment simulates a classroom, it could gauge actual teaching behavior. Individuals are likely to respond to instruction differently even when exposed to similar content. The simulation-based classroom standardizes the classroom environment, allowing for more equitable comparisons across educators. Such simulation-based effectiveness measures (SBEMs) would take advantage of innovations in instructional technology and in the field of educational measurement to better capture high-quality teaching.

First, I discuss the challenges in measuring teacher effectiveness for all teachers, and for teachers of the gifted in particular, as these teachers are susceptible to unfair evaluations because of the population they serve. Then I describe how SBEMs could be developed and how SBEMs could counterbalance these weaknesses if they were incorporated into an effectiveness rating system.

## Challenges With Measuring Teacher Effectiveness in Gifted Education

One challenge with measuring teacher effectiveness is that the definition of good teaching changes depending on the content addressed and students served. Teachers deemed to be effective with one group may not be as effective with others (Brophy & Good, 1986), requiring effectiveness ratings to be conditioned on the characteristics of students and/or content areas involved. Therefore, specialized measures are likely needed.

With respect to gifted students, much is known about effective instructional practice (Reis & Renzulli, 2011; Robinson, Shore, & Enerson, 2007; Tomlinson, 2005) and therefore implementation of strategies that are widely believed to be effective can be observed. However, effective teaching involves more than using a good strategy, it requires knowing when a particular approach will be most effective and how to meaningfully implement that approach for a given group of students studying a specific topic.

Therefore, knowledge of the students served and, of course, content is needed to accurately judge teacher effectiveness. However, it is too expensive and logistically difficult to implement a measure of teacher effectiveness that adequately accounts for context on a large-scale basis. A common and growing approach is to examine the value a teacher adds to achievement by comparing prior achievement and the achievement attained at the end of the instruction using any of a wide variety of methods (McCaffrey, Koretz, Lockwood, & Hamilton, 2003). This approach assumes that scores will improve only if teachers have done a good job within their classroom context.

Although tests scores are one important indicator of teacher effectiveness, the tests that tend to be used in these analyses are often designed to capture the skills of average students and are less helpful in describing the gains of gifted learners for several reasons. First, test scores are not uniformly reliable. Scores contain more error (and are therefore less accurate) for very low– and very high–performing students than they are for average students (Lohman & Korb, 2006). Tests are also limited in terms of the range of abilities they are designed to capture. Many gifted students may master the tested content but then vary in how much they know with regard to more advanced skills. As they have uniformly mastered tested abilities, student performance does not differ because there is a ceiling on the amount of knowledge the test can capture (McBee, 2010). Therefore, students may receive similar growth estimates, even though the amount that they have learned differs. In addition, the amount of gain that can be captured for initially high-performing students will be smaller than for those who started out with lower scores, making their teachers appear less effective. Statistical artifacts (e.g., regression to the mean) also make it highly likely that the test score gains of gifted students will be smaller than those of lower achieving students (Campbell & Kenny, 1999).

## A New Approach

Current methods provide valuable information about different aspects of teacher effectiveness. However, they are flawed in that they either provide indirect insight into what teachers actually do in the classroom, are not available for all teachers, or are dependent on the students served. Therefore, they cannot be used to ascertain how effective teachers might be if they worked with different students or to infer that teachers with higher effectiveness ratings are better teachers than those with lower scores. Most users of effectiveness measures do just this. They make employment decisions, award merit, and even assign certification based on these results. A measure of effectiveness that puts teachers on an even playing field is needed. To achieve this, the measure should be based on teachers' interactions with a set group students, students learning the same content and who exhibit the same classroom behaviors. By capitalizing on innovations in two fields—educational measurement and technology—we can develop SBEMs to accomplish this goal.

SBEMs would gauge effectiveness based on teachers' work in virtual reality classrooms. Teachers complete the SBEM after having time to review information about the classroom, including the particulars of the students and after preparing teaching materials for a lesson with a predetermined duration. Teachers would then dress in virtual reality garb—put on a headset with speakers to hear student speech and a microphone to record their own comments, and hold onto or attach a motion detector like a Nintendo Wii® remote onto their arms—and enter the virtual reality classroom, a room with a whiteboard that they can write on directly or use as a screen for projecting images, placed opposite a screen on which they can view their avatar students.

Students would be programmed with different personalities, capabilities, and behaviors. Wiggly Annie might be likely to fall off her chair when too bored or excited, "scowly" Kyle might give the teacher death stares and attempt to disrupt the class by taking other students' pencils and by making paper airplanes, and "too smart" Sally might raise her hand constantly and attempt to answer every question. Students would also exhibit common challenges in learning material—problems organizing work, forgetting what step they are on in solving three-digit multiplication problems, or have difficulty reading a word problem. The teachers would provide instruction just as they would in a regular classroom and would be able to see their own avatar on the screen and to move between students as needed. The virtual reality gear and the whiteboard would be used to capture how teachers responded to students and what they did and said in providing instruction. Avatar students would respond to examinees in real time based on the teachers' movements and speech, similar to the way that video games interact with users. Moreover, similar to a video game, teachers' actions would be scored using computer algorithms in real time.

One strength of this method is that SBEMs can create situations designed to elicit specific teaching behaviors. For example, assessments could be created specifically to assess a teacher's skill in working with gifted students. Teachers might be placed in a virtual classroom that included gifted, average, and cognitively disabled third-grade students and asked to implement a lesson on double-digit multiplication. The SBEM could capture whether the teacher differentiated instruction to meet the needs of the disabled, average, and gifted students; the differentiation methods used; and the effectiveness of teachers' attempts to scaffold the material. One advantage to this approach is that teachers would receive feedback both in the form of test scores and from student reactions to the teachers' efforts. Students might get bored and act out if the material were too easy or break into tears if too difficult. They also would communicate with the teacher about their misconceptions, allowing SBEM developers to gauge how well teachers respond to student questions. Finally, this method allays fears that evaluation results are inflated or depressed due to student characteristics because teachers are all exposed to the same classroom and the same students.

The principal–teacher conference presented at the beginning of this article would not change drastically if SBEMs were included among the effectiveness measures, but it would provide a measure that could be directly compared across a set of teachers. We still must hold teachers accountable for the academic achievement of their students, and we still expect principals to observe classrooms and to give feedback to their teachers. However, adding the evidence generated by SBEMs may allay fears that evaluations are too subjective and not based enough on comparable measures by including a standardized assessment. SBEMs also provide a direct measure of teaching, which may be more acceptable to teachers. Especially when different pieces of evidence appear to present contradicting results, additional information that is fair, comparable, and that directly measures the construct of interest may help to inform stakeholders.

Despite the need for SBEMs, there are obviously many challenges in developing these assessments, including creating the technology required for such a measure, identifying a progression of teaching skills that can be used to place teachers on a continuum from novice to expert, determining what kinds of experiences would elicit information about teacher performance on that continuum, developing the assessment, and generating a scoring system that would accurately reflect teacher effectiveness across a range of grade levels and content areas. Current work in the fields of technology-based performance assessment and in the area of educational measurement referred to as evidence-centered assessment design (ECD) could inform SBEMs. These fields, and their contribution to SBEMs, are discussed below.

## *Technology-Based Performance Assessment*

Simulation-based assessments have been developed in a wide array of fields and can inform the development of SBEMs. The National Board of Medical Examiners includes a computer-based simulation in its U.S. Medical Licensing Exam (Federation of State Medical Boards of the United States & National Board of Medical Examiners, 2011) that requires future physicians to diagnose and treat a patient by reading patient's medical history, ordering tests and receiving results, and getting periodic updates about the patient's condition, which varies according to the tests and treatments administered. Cisco systems has also developed an interactive assessment, called Cisco Aspire®, which involves having information technology professionals negotiate networking projects, purchase equipment using the budget they negotiated, and identify and fix networking problems. As would occur with SBEMs, and is rare in current assessment efforts, both systems change the stimuli examinees experience based on their earlier responses to an extended, open-ended problem and provide real-time, computer-generated scoring. The exams also present several different simulations embedded within the assessment to gauge performance in a variety of situations on many different tasks, allowing scores to represent achievement of a wide variety of skills. This is also rare in traditional performance assessment due to the expense associated with hand-scoring.

Several assessments have also been developed in elementary and secondary education to gauge student proficiency in science. Virtual performance assessments (VPAs), developed by Chris Dede and colleagues at Harvard University, also involve open-ended tasks (Clarke-Midura, Code, Mayrath, & Dede, 2011). VPAs engage students in video games involving science inquiry (e.g., students explore a seaside area, interview people, and collect scientific data to determine why kelp are dying and to come up with a strategy to improve the environment). As with the Cisco and Graduate Management Admission Test examples, VPAs use computer-generated scoring; behavior throughout the task is monitored and contributes to the score.

In addition, the Situated Multimedia Arts Learning Laboratory (SMALLab) at Arizona State University is closest in implementation of my vision for SBEMs in that it involves placing people into a space in which they interact with their environment (Johnson-Glenberg, Birchfield, Savvides, & Megowan-Romanowicz, 2010). The

SMALLabs incorporate three dimensional object tracking, real-time graphics, and surround-sound to enhance the examinees' experience during assessment and to capture examinee movements and speech to help with scoring. Examinees get feedback on their efforts in real time during the assessment with visual displays that show the results of their work, similar to the way that teachers get feedback about the quality of their instruction from students as they teach. SMALLabs are designed with feedback in mind, with the philosophy that students will learn by completing this formative assessment. Many teacher-effectiveness efforts are also intended to help teachers improve in addition to holding them accountable for performance (Georgia Department of Education, 2011). SBEMs may actually improve instruction by helping teachers see their strengths and weaknesses during and immediately after the assessment when they can best recall exactly what actions lead to their score.

Finally, simulated classrooms have already been developed and are currently being used in teacher preparation programs to provide formative information. This work can greatly inform SBEM efforts. The TeachME program at the University of Central Florida involves exposing preservice teachers to simulated classrooms. Preservice teachers stand in front of a large screen on which they can view five avatar students while wearing a headset to speak with them. Actors are hired to play students in the classroom and respond to the teachers' efforts with guidance from faculty (Kolowich, 2010). SimSchool© is a computer-based game, much like the SimCity® game series. Preservice teachers sit at a computer and respond to a classroom full of avatars by typing instructions or making selections with a mouse. This is a less realistic setting but one in which actors are not required because all student interactions are programmed. My vision of SBEMs would merge these two efforts—incorporating the efficiency of programmed student–teacher interactions with the more realistic setting offered by a virtual classroom.

## ECD

Evidence-centered assessment design (ECD; Mislevy & Haertel, 2006) is a relatively new approach to educational measurement that is particularly well suited for performance assessment and could be used to inform the SBEM test development process. Its structure undergirds most of the technology-based performance assessments mentioned above. In contrast with other measurement approaches, which focus squarely on the statistical methodology involved in generating accurate scores, ECD provides a somewhat loose framework for test development that is intentionally nebulous so that it can be adapted to a variety of situations.

Despite this, the method focuses with laser-like intensity on providing evidence for or against mastery of the assessed construct. At its heart, ECD helps test developers to think deeply about what mastery would look like, what behaviors best present evidence of mastery, and how to construct an assessment that will elicit these behaviors, allow for accurate scoring, and at the same time maximize efficiency. It keeps the construct of interest at its center, and all decisions are made with concern for an assessment design that produces the best possible evidence to support a particular score. This is especially helpful with complex assessments like SBEMs as the overall goal of the assessment can be lost amid the myriad of decisions that must be made to get the measure up and running.

The basic framework of ECD involves five basic steps or layers: Domain Analysis, Domain Modeling, the Conceptual Assessment Framework, Assessment Implementation, and Assessment Delivery. Each step is discussed below in relation to SBEMs.

*Domain Analysis*. Under ECD, it is important to fully explicate the construct assessed and to create scenarios that will reveal the degree of construct attainment *before* attending to how tests and items are developed or scored. That is, the domain of interest is very thoroughly explored and defined using processes that are much more involved than is typical of traditional test development. In traditional test development, the domain is specified with a table of specifications, and many more items are generated than are actually needed for each cell of the table. Items are winnowed and revised with equal emphasis placed on item statistics and on test content (Hambleton & Jones, 1993).

In contrast, the first step in designing a SBEM is to go beyond determining the basics of the construct to be assessed—what grade levels, content areas, and student needs are most relevant—and also to decide which aspects of instruction should be explored. This also involves devoting many months of intensive study about what it means to be a good teacher. Test developers might first depend on learning theory and the existing literature and would follow these with empirical analyses.

One approach is detailing the progression of teaching skills from those evidenced by novice to expert teachers on a wide variety of skills that will be assessed—organization and management, interactions with students, ability to match instruction to student ability levels, clarity of explanations, and so on. To do so, verbal analysis is often used. Verbal analysis requires experts to "think aloud" about a topic, for example, to provide a list of all the strategies that they might use in presenting a multiplication problem, and then uses their responses (frequency with which approaches are mentioned and proximity of approaches in the conversation—e.g., teachers tend to mention direct instruction and independent practice concurrently and group activities separately) to develop a model of task performance (Ericsson & Simon, 1993; Leighton, 2009).

*Domain Modeling*. After the progression of skills has been developed, the next step is to decide what kinds of evidence would help position examinees along the progression. This step is a precursor to designing the assessment itself; it answers the question, "Now that I know what I want to assess, what types of things might happen to provide evidence to support assertions about teaching effectiveness?" In this step, test developers create scenarios about exactly what kinds of behaviors teachers would need to exhibit to place teachers at different points along the effectiveness continuum.

The data collected during the Domain Analysis stage are crucial in these efforts, as is the use of expert judgment to determine what the assessment will measure and how it will operationalize the skills of interest. In addition, great attention is given to developing the rationale behind each decision: why each behavior was selected and why it was associated with different positions along the continuum. Finding the right "grain size" is a key concern in this process. Progressions that too narrowly define each step do not allow for variation in a set of behaviors that might all represent equivalent levels of effectiveness, while definitions that are too broad provide so little information about the differences between teachers at different steps that they are not helpful (Leighton, 2009).

*Conceptual Assessment Framework*. Test developers only move onto the Conceptual Assessment Framework phase after the Domain Analysis and Domain Modeling steps are complete. In this stage, test developers create technical specifications that lay out assessment details such as the general framework used in scoring, the measurement model used (e.g., Item Response Theory, Generalizability Theory, Bayesian analysis), and the most appropriate way to administer the test (e.g., Are simulations really needed or would a more straightforward method suffice? What technological innovations should be used to best get at the constructs of interest?). As in the previous stages, all decisions are made carefully, and the reasons behind each choice should clearly indicate that the selected methods provide the best evidence to inform scoring. This stage is especially important for SBEMs because teaching is such a complex task that decisions must be made about which skills are best gauged in a technology-based simulation and which can be addressed through different data collection efforts. When teaching efforts can be measured simply, they should be. The simulations should focus on aspects of teaching that cannot be measured well through other means.

*Assessment Implementation*. Assessment Implementation is concerned with creating the operational test—writing tasks or items, finalizing scoring systems, producing test forms or the computer algorithms that will guide test administration, and estimating parameters for measurement models. For SBEMs, this is also the stage at which the virtual reality environment will be developed and tested. In many ways, this is quite similar to other test development efforts. ECD is unique in that decisions continue to be made by placing the construct of interest at the center. For example, test developers will have to make a myriad of decisions about how teachers will experience the virtual reality environment. As virtual reality is by definition artificial, and unlike a regular classroom, developers will need to consider which aspects of classroom life must be quite similar to the real world and which can be dissimilar. Under ECD, each decision will be made based on how it will affect test score interpretations with concerns about expense and ease of administration still important but secondary.

*Assessment Delivery*. Finally, the assessment is administered and scored in the Assessment Delivery layer. This is also quite similar to conventional testing practice in that it is concerned with the operational aspects of test administration, including generating scores and score reports. It also bears some similarity to computer-adaptive testing; because SBEMs are fluid, with student avatars responding to teacher's efforts, a large number of items must be created and placed in a library with teachers exposed to a subset of items based on their characteristics and the responses they give during the assessment. Procedures for determining which examinees see which items, the order in which items are presented, and processes for scoring the assessment would also be finalized in this phase. ECD purports to be unique in that it creates systems that optimize efficiency by ensuring that "items" in the form of student behaviors are easily interchanged depending on the simulation's needs and can be reused in different situations.

This brief discussion of ECD is intended to impart some basic information about how SBEMs might be implemented. Mislevy and Riconscente (2006) provided a far more detailed greater explanation of ECD and is an excellent resource.

Next, I examine methods currently used to measure teacher effectiveness and discuss how SBEMs could improve accountability systems concerned with teacher effectiveness.

## Current Measures of Teacher Effectiveness

Three general approaches to capturing teacher effectiveness dominate the literature: (a) classroom observations, (b) student achievement measures, and (c) teacher tests. Different researchers implement these approaches in different ways, depending on the content or grade level addressed and the beliefs of the researcher.

All three approaches have serious limitations. Observation measures are confounded with the students taught and can only tell us about a teacher's accomplishments with a given group of students. However, when used for accountability, we assume that scores are comparable so that Miss Jones's instruction in a kindergarten classroom can be scored on the same metric as Mr. Shang's teaching in a high school physics class. This is important because teachers can only be fairly evaluated for accountability purposes when scores are comparable.

Student achievement measures of effectiveness assume that test score improvement is due to classroom instruction instead of experiences outside of school and that the tests used are sensitive to instructional efforts and do not simply reflect student aptitude. In addition, to be universally applicable, students must also be tested in all subject areas and at every grade level, which is not current practice.

Finally, although teacher certification exams are not dependent on student characteristics, they can only gauge academic understanding of teaching skill and of content knowledge rather than how that understanding is applied. This is a serious limitation in that conceptual understanding and ability to implement that understanding are distinct skills. Each approach is discussed in detail below.

### *Classroom Observation*

Classroom observations gauge teaching effectiveness by rating characteristics of instruction observed during a finite period of time. Therefore, observations provide the most direct method of evaluation in that ratings are based on teacher–student interactions and do not require the assumption that improvements in student achievement or teacher test performance are valid indicators of classroom interactions. As they are more proximal to instruction, classroom observations are widely considered to be the most accepted measures of effectiveness with educators (Heneman, Milanowski, Kimball, & Odden, 2006).

This approach assumes that, if teachers are effective, they will implement the desirable behaviors during the observation. It also assumes that universal characteristics of effective instruction can be observed; will permeate instruction across content areas, students served, and time; and can be captured through a standardized observation protocol. Milanowski (2011), Pianta and Hamre (2009), and Danielson and McGreal (2000), among others, had argued that classroom observations are essential in gauging teacher effectiveness. However, they also acknowledge that these measures are limited in terms of the aspects of effectiveness that can be assessed. As Danielson and McGreal (2000) explained,

> Classroom observation is a critical evaluation methodology for those aspects of teaching that may be directly observed. Some important aspects of instruction, however—even those involving a teacher's work with students, such as providing feedback to students—are not necessarily easily observed in a classroom episode . . . Similarly, a teacher's skill in establishing classroom routines may not be observed directly, but rather inferred from the behavior of students as they go about their business, seemingly with no direction from the teacher. (p. 47)

When observable characteristics are of interest, several concerns remain. There is a great deal of potential for bias. First, teachers may change their behavior because they are being observed, resulting in observations that do not reflect typical instruction, a phenomenon commonly referred to as a Hawthorne effect (Adair, 1984). Second, the raters themselves may unwittingly provide biased scores. Halo effects, or the extent to which global perceptions of a teacher affect all ratings, are one form of bias. For example, a rater might have a favorable impression of a teacher who is well organized and caring with students. If that teacher then erroneously explains a concept, the rater might score the accuracy of explanations higher than he or she might for a teacher with an equally poor explanation but a less caring affect. This problem occurs even when the observer knows the teacher well and should therefore be familiar with the individual's strength and weaknesses.

In their study of the ratings assigned to 161 student teachers by their cooperating teachers, Phelps, Schmitz, and Boatright (1986) found highly consistent ratings on five different characteristics—attitude, scholarship, instruction, discipline, and personality. Ratings were so consistent that instruction and personality in combination accounted for variation in ratings equally well as the combination of all five factors. In addition, cooperating teachers were lenient in their ratings, assigning scores that Phelps et al. considered to be spuriously inflated because scholarship ratings were not consistent with student teachers' grade point average (GPA), American College Testing (ACT) score, or grade in a learning and instruction class.

Mujis (2006) recommended several steps to improve the quality of classroom observation ratings. First, to improve interrater reliability, or the likelihood that two people would assign the same rating to an observation, it is helpful to use low-inference indicators, which require a small amount of judgment, whenever possible. However, it is more important to ensure that indicators best reflect effective teaching, which likely involves a higher degree of inference. Second, rater training can decrease the amount of bias in scores, reducing both the halo effect and other rater effects that occur because some tend to be overly lenient or overly rigorous in assigning scores. Standardizing observation protocols ensures that raters are systematic and helps to ensure the reliability and validity of scores (Pianta & Hamre, 2009). Milanowski (2011) also suggested that raters collect instructional artifacts to assess abilities in nonobservable competencies like planning and assessment. Finally, as Chism (2007) and others pointed out, multiple observers and multiple observations are required to ensure accurate scoring.

These steps will improve the quality of ratings based on observations, a critical method in gauging teacher effectiveness. However, one issue with conducing classroom observations remains. As they are resource intensive, a limited number of observations can be conducted per classroom. For example, Georgia's state teacher evaluation system requires a minimum of two informal observations, lasting 5 to 15 minutes each, and one 30 to 50 minutes formal observation (Georgia Department of Education, 2011). In that state, teachers must be observed 3 times but could be observed for only 40 minutes total across the three sessions. It is unlikely that observation scores will generalize across content areas, students, and time under this scenario because an elementary teacher may be highly effective at teaching reading but less so at teaching mathematics. If observations happen to occur during reading time, or if observations of instruction in different content areas are extremely brief, effectiveness ratings may be inflated.

Similarly, a teacher may be more effective in their interactions with particular kinds of students or may be more effective at certain points in the school year (Mujis, 2006). However, those who interpret the scores are likely to assume that they universally apply. Finally, although several classroom observation instruments have been developed specifically for classrooms of gifted students, including the Classroom Practices Record (Westberg, Archambault, Dobyns, & Salvin, 1993), the Teacher Observation Form (Peters & Gates, 2010), and the Classroom Observation Scale–Revised (VanTassel-Baska et al., 2005), I could not find an instance of such measures being used for accountability purposes.

SBEMs address these concerns in several ways. First, because they are computer scored, rater bias is not a concern. Albeit, scoring algorithms will have to be carefully constructed and could lead to other forms of error if not properly implemented. Second, because test developers control the virtual classroom, they can create environments that elicit the skills we want to measure. If classroom management is of key concern, the students can be a little rowdy and then either amp up or tone down their behavior based on teacher efforts. SBEMs can also test teacher pedagogical content knowledge (PCK) by looking for common mistakes that teachers make in explaining content and their facility in addressing common misconceptions among students.

### *Student Achievement Measures*

When teacher effectiveness is defined in terms of improvement made during the school year, students' test performance adjusted for prior achievement is the main indicator of effectiveness. Many statistical techniques can be used to gauge effectiveness using test scores, most of which fall under a general approach called value-added modeling (McCaffrey et al., 2003; Sanders, Wright, & Horn, 1997).

Value-added models are a prevalent indicator of effectiveness and are appealing in that they reflect the desired outcome of education, student learning, instead of educational processes. In fact, those who call for improved measures of teacher effectiveness often insist that value-added scores are included (Gordon, Kane, & Staiger, 2009). Value-added models also greatly improve on previous accountability efforts, which examined student performance without regard for what students knew at the beginning of the school year (No Child Left Behind Act, 2002).

In contrast, value-added models examine test performance at the end of the school year after adjusting for prior performance. There are many ways to accomplish this. For example, some models assume all gains are due to the current teacher, while others allow for teacher effects to persist over time, measuring the effects of the third-, fourth-, and fifth-grade teachers on gains made in fifth grade. Moreover, when teacher impacts are allowed to persist, the ways in which they are allowed to persist varies (e.g., the third-grade teacher is weighted equally with the fifth-grade teacher or their effect diminishes over time). Models also vary in terms of the student characteristics included. While some approaches adjust for demographic characteristics such as ethnicity and socioeconomic status, others only account for prior achievement (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). It should be no surprise then that estimates of teacher effectiveness vary depending on the method used, resulting in different rank orders of teachers (Lockwood, Louis, & McCaffrey, 2002; McCaffrey et al., 2004).

In addition, value-added results have been found to vary by test, with teachers ranked differently on different achievement measures (Papay, 2011). Varied results might occur because some tests are limited in their ability to detect instructional effects. For value-added models to work, test scores must be sensitive to instruction—students whose instructor teachers test content well should answer items correctly and those whose instructor does not cover test content, or does so badly, should provide incorrect responses. When tests scores improve as a result of instruction, the test is instructionally sensitive (Domaleski & Hill, 2010; Popham, 2007). Several approaches to measuring instructional sensitivity have been proposed, including using a pretest–posttest design to determine whether scores improve after instruction on tested content

(Popham, 2007), but few instructional sensitivity studies have been conducted (Polikoff, 2010). The studies that do exist have found that tests are limited in the degree to which they reflect instructional efforts (D'Agostino, Welsh, & Corson, 2007), and value-added analysis would therefore yield spurious results. Of equal concern is that few states conduct instructional sensitivity analysis at all, yet implement high-stakes accountability systems without establishing how well scores reflect instructional efforts. This has lead Baker (2008) and others (Polikoff, 2010) to suggest that instructional sensitivity should be central to the test-validation process.

Instructional sensitivity may be of even greater import for gifted students, who are likely to exhibit high test scores regardless of instructional efforts, limiting the magnitude of gains that are possible to observe. In addition, if gifted students receive specialized instruction outside of school, the effects of classroom and outside experiences will be confounded. Assuming that gifted students are more likely than others to participate in academic training beyond the school day, these factors make value-added modeling particularly complex. Empirical work supports this concern. Rambo (2011) examined the learning gains made during the school year and over the summer in 2,000 schools nationwide on a computer-adaptive test. She found that gifted students gained at a consistent rate during the summer and during the school year in reading, in contrast with average students who experienced school-year gains and summer loss in achievement.

Finally, value-added modeling requires student test scores at the beginning and the end of the current school year (or in the spring of two consecutive years). As most teachers teach untested subjects like art, social studies, physical education, and science (which have state tests at a limited number of grade levels) or at untested grade levels, value-added models address the effectiveness of a minority of teachers (Prince et al., 2009). Methods for measuring the contribution to student growth made by teachers in nontested grades and subjects are under investigation (Meyer, 2010), but no strategy has emerged as the superior approach (Goe & Holdheide, 2011).

Although they should not be used to supplant value-added models, SBEMs could provide supplementary evidence of teacher effectiveness within a multiple measure framework and could be particularly useful in measuring the effectiveness of teachers of the gifted because they can be focused to address those components of a lesson most relevant to gifted students (e.g., Did the teacher explain things or develop activities that were at the appropriate level for gifted students and that clearly transmitted key information? Did they assess the right skills and respond appropriately to assessment results? Did they use open-ended activities?). As the achievement of gifted students is likely to be very high, and may confound in-school and out-of-school educational experiences, value-added models might be less informative about teachers of the gifted necessitating additional effectiveness measures. In addition, SBEMs could be created to address the grade levels and content areas overlooked by student assessments and administered with less expense as there are far fewer examinees at the teacher than the student level.

## Teacher Tests

When effectiveness is measured through teacher tests, it is operationalized as knowledge of instructional content and of methods to convey that content. These tests typically assess whether teachers possess the minimum level of content knowledge required for effective instruction and also address approaches to presenting material and other pedagogical issues. As they incorporate both knowledge of content and knowledge of teaching, these tests are often referred to as measures of PCK (Shulman, 1987).

Kromrey and Renfrow (1991) are often cited as early scholars of PCK. Using Shulman's (1987) explanation of the difference between content knowledge, pedagogical knowledge, and PCK, they propose a set of general skills that might be assessed by a test of PCK, including error diagnosis, communicating with students, organizing instruction, and learner characteristics. Although these broad topics suggest an emphasis on pedagogy instead of content, Kromrey and Renfrow (1991) emphasized that items must be situated within a specific skill set using scenarios or situations for examinees to consider. For example, in reference to the skill "evaluate student homework," they provide the following example:

> Which feedback is most appropriate for a six-year-old first grader who wrote a story about "nites in shng armr ftng dragnz?"; how should a teacher provide feedback regarding a student's customer letter responding to a delayed order for a business communication class? (p. 7)

They also explain that PCK items require an awareness of the teaching process instead of content knowledge alone, "Items reflect the process of teaching the content, not the noninstructional practice of the discipline" (Kromrey & Renfrow, 1991, p. 5).

The concept of PCK has permeated teacher certification exams, with pencil-and-paper exams commonly required for teaching licensure and also part of the process for becoming certified by the National Board of Professional Teaching Standards, an elite status for which teachers often receive merit pay (Ball, Thames, & Phelps, 2008; Educational Testing Service, 2006; Rowan, Schilling, Ball, & Miller, 2001). However, questions have been raised about the ability of these tests to capture teaching skill. In their meta-analysis of 123 studies, D'Agostino and Powers (2009) found that GPA in teacher preparation program was a better predictor of teaching competence than teacher certification exams. In addition, Goldhaber and Hansen (2010) examined the certification exam used in North Carolina over a 10-year period and found that the tests were differentially predictive of teacher impact on student performance, with the tests being poorer predictors for African American and male teachers; African American teachers who performed worst on the certification exam appeared to be the most effective at improving the academic achievement of disadvantaged students.

National Board Certification (NBC) has also come under some scrutiny. Goldhaber and Anthony (2007) found that the NBC process succeeds at identifying the more competent teachers among all NBC applicants but that more effective teachers also self-select by applying for NBC certification. However, the strength of the relationship between student achievement and NBC certification varied by grade level and by student characteristics. In contrast, Stronge et al. (2007) compared NBC teachers and their noncertified counterparts in four North Carolina school districts on a variety of characteristics and did not find much of an NBC effect. Although the NBC teachers outperformed their counterparts in terms of the clarity of their assignments, the level of cognitive complexity required by assignments, and planning practices, they were indistinguishable from other teachers in terms of behaviors observed during instruction and in terms of student achievement.

Measures of PCK have also been developed as research instruments. Heather Hill and colleagues (2008) had developed a paper-and-pencil test of Mathematical Knowledge for Teaching (MKT), which is akin to PCK, and an observational tool relating directly to the quality of mathematics instruction. After using both measures with the same group of teachers, they compared scores and concluded that MKT is associated with instructional quality. More recently, Hill, Kapitula, and Umland (2011) compared both MKT and instructional quality with value-added test scores and found all three measures to be intercorrelated, a somewhat different outcome than observed in the NBC studies. The MKT results are based on small and somewhat specialized samples, of ten teachers and 24 teachers, respectively, each drawn from one school district. More study is needed to confirm that the measures reflect quality of instruction in a variety of settings.

There are practical concerns associated with using teacher tests to gauge teacher effectiveness that extend beyond the belief that test performance may not adequately reflect classroom activities. Few states offer teacher certification exams in gifted education. Therefore, the infrastructure does not currently exist to gauge teacher effectiveness for this group. In addition, although many characteristics of good teaching transcend the types of students involved, the needs of gifted students are unique and require specialized instruction. In their investigation of the factors associated with the underachievement of gifted students, Reis and McCoach (2000) identified boredom with the regular curriculum among the factors that contribute to underachievement. Strategies that better challenge gifted students abound (e.g., differentiation, curriculum compacting, use of open-ended activities) and could be addressed on a test for teachers of the gifted.

SBEMs improve on paper-and-pencil tests of teacher knowledge because they are performance based. Tests of PCK improve on previous measures in that they measure knowledge of the challenges associated with teaching course content. Even so, knowledge of the challenges that occur in teaching, and of the best way to present information, is only a precursor at best to effective teaching: It gauges whether a teacher knows what to look for and what to do but does not directly assess what the teacher does once placed in a classroom. SBEMs provide a more direct, if somewhat artificial, measure of how teachers operationalize their knowledge.

## Limitations of SBEMs

I would be remiss if I did not acknowledge that SBEMs also suffer from their own limitations. One key concern is that SBEMs will inevitably be artificial and therefore unlike real classrooms in many ways. Although they are more direct measures of teaching effectiveness than teacher tests and student achievement measures, they are not particularly authentic assessments of teaching skill. This is likely to be a greater concern for some aspects of effectiveness than for others. For example, teachers may not have the same interactions with avatars as they would with real students. As affective aspects of interaction often transcend a particular content area, they may be better assessed in classroom observations.

Another concern is that teachers may vary in their ability to operate the equipment used in the simulation such that teacher-effectiveness ratings are confounded with ability to operate the technology. Although my vision for SBEMs uses relatively widespread and straightforward technology (headsets and video game remotes), this is certainly a concern. Although teachers would have opportunities to practice with the equipment before participating in the assessment, this may not be adequate. The effect of technical capacity on scores is one of many concerns requiring further study before SBEMs could be implemented on a large-scale basis.

## Conclusion

Teacher effectiveness is a burgeoning field. Policy makers seem increasingly committed to the concept that teacher effectiveness should be gauged on a large-scale basis and that individual teachers should be held accountable for their effectiveness ratings. Despite their strengths, it is clear that current measures could easily be misused. Scores must be generalizable across classrooms and content areas for use in accountability. However, because teachers work in a very specific context, with a particular group of students, it is impossible to know what their scores might look like if evaluated in a different classroom.

This is of particular concern in gifted education, where interactions with gifted students will be ignored or subsumed with those of average students in integrated settings where gifted students tend to be a small minority. And there is some chance that teachers of the gifted will look worse than other teachers when they work in specialized classrooms due to statistical and measurement artifacts, especially when based on student achievement gains, an increasingly popular technique.

Simulated effectiveness measures overcome these weaknesses by measuring effectiveness directly, based on performance in a classroom setting, and by standardizing both the students and content taught, which allows for direct comparisons across teachers. Innovations in technology and in assessment make it possible for us to do a much better and a much fairer job of evaluating teachers. There is still much work to be done in this area. However, this article attempts to shine a light on what is possible and to energize the field to improve on current practice.

## Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author received no financial support for the research, authorship, and/or publication of this article.

## References

Adair, G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology, 69*, 334-345.

American Recovery and Reinvestment Act of 2009: P. L. 111-5, as signed by the President on February 17, 2009: Law, explanation, and analysis. (2009). Chicago, IL: CCH.

Baker, E. L. (2008, April). *Empirically determining the instructional sensitivity of an accountability test*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*, 389-407.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Whitrock (Ed.), *Handbook of research on teaching* (pp. 328-375). New York, NY: Macmillan.

Campbell, D. T., & Kenny, D. A. (1999). *A primer of regression artifacts*. New York, NY: Guilford.

Campbell, R. J., Kyriakides, L., Mujis, R. D., & Robinson, W. (2004). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education, 29*, 347-362.

Chism, N. V. N. (2007). *Peer review of teaching: A sourcebook* (2nd ed.). Bolton, MA: Anker.

Clarke-Midura, J., Code, J., Mayrath, M., & Dede, C. (2011, April). *Exploring inquiry processes in virtual performance assessments*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

D'Agostino, J. V., & Powers, S. J. (2009). Predicting teacher performance with test scores and grade point average: A meta-analysis. *American Educational Research Journal, 46*, 146-182.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional validity of a state's standards-based assessment. *Educational Assessment, 12*, 1-22.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Princeton, NJ: Association for Supervision and Curriculum Development.

Dixon, A. (2011). *Focus on teacher reform legislation in SREB states: Evaluation policies*. Atlanta, GA: Southern Regional Educational Board. Retrieved from http://publications.sreb.org/2011/11S07_Focus_Teach_Eval.pdf

Domaleski, C., & Hill, R. (2010). *Considerations for using assessment data to inform determinations of teacher effectiveness*. Dover, NH: National Center for Improving Educational Assessment. Retrieved from http://www.nciea.org/papers-UsingAssessmentData4-29-10.pdf

Educational Testing Service. (2006). *Proper use of the Praxis series and related assessments*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/PRAXIS/pdf/guidelines.pdf

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. Cambridge, MA: MIT Press.

Federation of State Medical Boards of the United States & National Board of Medical Examiners. (2011). *United States Medical Licensing Exam: 2011 Bulletin of information*. Philadelphia, PA: Author. Retrieved from http://www.usmle.org/General_Information/bulletin/2011/2011%20BOI.pdf.

Georgia Department of Education. (2011). *CLASS keys: Classroom analysis of state standards: The Georgia teacher evaluation system*. Atlanta, GA: Author. Retrieved from http://www.gadoe.org/DMGetDocument.aspx/CK%20Standards%2010-18-2010.pdf?p=6CC6799F8C1371F6B59CF81E4ECD54E63F615CF1D9441A92E28BFA2A0AB27E3E&Type=D

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Princeton, NJ: Educational Testing Service.

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for nontested grades and subjects*. Washington, DC: National Comprehensive Center on Teacher Quality.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *Review of Economics and Statistics, 89*, 134-150.

Goldhaber, D., & Hansen, M. (2010). Race, gender, and teacher testing: How objective a tool is teacher licensure testing? *American Educational Research Journal, 47*, 218-251.

Gordon, R., Kane, T. J., & Staiger, D. O. (2009). *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institution.

Hambleton, R. H., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.

Heneman, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge and skill-based pay* (CPRE Research Report RB-45). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education. Retrieved from http://www.cpre.org/images/stories/cpre_pdfs/RB45.pdf

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430-511.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*, 794-831.

Johnson-Glenberg, M. C., Birchfield, D., Savvides, P., & Megowan-Romanowicz, C. (2010). Semi-virtual embodied learning-real world STEM assessment. In L. Annetta & S. Bronack (Eds.), *Serious educational game assessment: Practical methods and models for educational games, simulations and virtual worlds* (pp. 225-241). Rotterdam, Netherlands: Sense.

Kolowich, S. (2010, July 7). Avatars to teach the teachers. *Inside Higher Ed.* Retrieved from http://www.insidehighered.com/news/2010/07/07/avatars

Kromrey, J., & Renfrow, D. (1991, February). *Using multiple choice examination items to measure teachers' content-specific pedagogical knowledge*. Paper presented at the annual meeting of the Eastern Educational Research Association, Boston, MA.

Leighton, J. P. (2009, April). *Two types of think aloud interviews for educational measurement: Protocol and verbal analysis.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*, 255-270.

Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted, 29*, 451-484.

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly, 54*, 314-320.

McCaffrey, D. F., Koretz, D. M., Lockwood, J. R., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: Rand.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*, 67-101.

Meyer, R. H. (2010, December). *Value-added systems, accountability, and performance management in education: Promises and pitfalls*. Paper presented at the 2010 Annual Conference of the Iowa Educational Research and Evaluation Association, Cedar Falls.

Milanowski, A. (2011). Strategic measures of teacher performance. *Phi Delta Kappan, 92*, 19-25.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*, 6-20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Mujis, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*, 53-74.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425. (2002).

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*, 163-193.

Peters, S. J., & Gates, J. C. (2010). The teacher observation form: Revisions and updates. G*ifted Child Quarterly, 54*, 179-188.

Phelps, L., Schmitz, C. D., & Boatright, B. (1986). The effects of halo and leniency on cooperating teacher reports using Likert-type rating scales. *Journal of Educational Research, 79*, 151-154.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*, 109-119.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*, 3-14.

Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89*, 146-155.

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009). *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades*. Washington, DC: Center for Educator Compensation Reform. Retrieved from http://www.cecr.ed.gov/guides/other69Percent.pdf

Rambo, K. E. (2011). *How much do schools matter? Using summer growth patterns to assess the impact of schools on high achieving and gifted students* (Unpublished doctoral dissertation). University of Connecticut, Storrs.

Reis, S. M., & McCoach, D. B. (2000). The underachievement of gifted students: What do we know and where do we go? *Gifted Child Quarterly, 44*, 152-170.

Reis, S. M., & Renzulli, J. S. (2011). Challenging gifted and talented learners with a continuum of research-based intervention strategies. In T. J. Kehle, M. A. Bray, & P. E. Nathan (Eds.), *Oxford handbook of school psychology* (pp. 456-482). Oxford, UK: Oxford University Press.

Robinson, A., Shore, B., & Enerson, D. (2007). *Best practices in gifted education: An evidence-based guide*. Waco, TX: Prufrock.

Rowan, B., Schilling, S. G., Ball, D. L., & Miller, R. (2001). *Measuring teachers' pedagogical content knowledge in surveys: An exploratory study*. Ann Arbor: University of Michigan.

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Shulman, L. S. (1987). Knowledge and teaching: Foundation of the new reform. *Harvard Educational Review, 57*, 1-22.

Stronge, J. H., Ward, T. J., Tucker, P. D., Hindman, J. L., McColsky, W., & Howard, B. (2007). National Board certified teachers and non-National Board certified teachers: Is there a difference in teacher effectiveness and student achievement? *Journal of Personnel Evaluation in Education, 20*, 185-210.

Tomlinson, C. (2005). Quality curriculum and instruction for highly able students. *Theory Into Practice, 44*, 160-166.

VanTassel-Baska, J., Avery, L., Struck, J., Feng, A., Bracken, B. A., Drummond, D., . . . Quek, C. (2005). *Classroom Observation Scale–Revised: User's manual*. Williamsburg, VA: The Center for Gifted Education, College of William and Mary.

Westberg, K. L., Archambault, F. X., Jr., Dobyns, S. M., & Salvin, T. (1993). *An observational study of instructional and curricular practices used with gifted and talented students in regular classrooms* (Research Monograph 93104). Storrs: The National Research Center on the Gifted and Talented, University of Connecticut.

## About the Author

**Megan E. Welsh** is an assistant professor in the Measurement, Evaluation, and Assessment Program in the Neag School of Education at the University of Connecticut. Her primary areas of research interest involve assessment, test validity, and the impact of these on educational reform efforts.