Do We Know a Successful Teacher When We See One? Experiments in the Identification of Effective Teachers

Journal of Teacher Education
62(4) 367–382

DOI: 10.1177/0022487110390221
http://jte.sagepub.com
SAGE

Michael Strong[1], John Gargani[2], and Özge Hacifazlioğlu[3]

Abstract

The authors report on three experiments designed to (a) test under increasingly more favorable conditions whether judges can correctly rate teachers of known ability to raise student achievement, (b) inquire about what criteria judges use when making their evaluations, and (c) determine which criteria are most predictive of a teacher's effectiveness. All three experiments resulted in high agreement among judges but low ability to identify effective teachers. Certain items on the established measure that are related to instructional behavior did reliably predict teacher effectiveness. The authors conclude that (a) judges, no matter how experienced, are unable to identify successful teachers; (b) certain cognitive operations may be contributing to this outcome; (c) it is desirable and possible to develop a new measure that does produce accurate predictions of a teacher's ability to raise student achievement test scores.

Keywords

teacher effectiveness, teacher evaluation, classroom observation, value-added

Since the No Child Left Behind Act (2002) became law, the term *teacher quality* has been close to the surface of many an educator's consciousness. Now, with President Obama's Race to the Top, there is a focus on teacher effectiveness. It is fairly well documented that the best school predictor of student outcomes is high-quality, effective teaching as defined by performance in the classroom (Goldhaber, 2002; Goldhaber & Brewer, 2000; Hanushek, Kain, O'Brien, & Rivkin, 2005; Wright, Horn, & Sanders, 1997). A high-quality teacher may have considerable impact on student learning. For example, Hanushek (1992) found that, all things being equal, a student with a very high-quality teacher will achieve a learning gain of 1.5 grade-level equivalents, whereas a student with a low-quality teacher achieves a gain of only 0.5 grade-level equivalents. This translates to one year's growth being attributable to teacher quality differences. More recently Aaronson, Barrow, and Sander (2007) examined data from the Chicago Public Schools and found that a one-standard-deviation, one-semester improvement in math teacher quality raised student math scores by 0.13 grade equivalents or, over one year, roughly one-fifth of average yearly gains.

William Sanders, who pioneered the Tennessee Value-Added Assessment System, summarizing his own studies, stated that especially in math, the cumulative and residual effects of teachers are still measurable at least four years after students leave a classroom (Sanders, 2000, p. 335). A study by Nye, Konstantopoulos, and Hedges (2004), unusual because it randomly assigned students to classes, estimated teacher effects on student achievement over four years. Their estimates of teacher effects on achievement gains were similar in magnitude to those of previous studies done by economists, but they found larger effects on mathematics than on reading achievement.

Observational Measurement of Teaching Practice

Findings such as these are convincing as to the importance of having an effective teacher but do nothing to tell us how to identify an effective teacher when we see one. Over the past few decades, researchers have attempted many ways of accomplishing this task, using a wide variety of first impressionistic and later systematic methods to investigate teaching practices through classroom observations (Brophy & Good, 1986; Nuthall & Alton-Lee, 1990; Stallings & Mohlman, 1988; Waxman, 1995), and findings from their studies have contributed to educators' notions of what constitutes good teaching. The several hundred observational systems that have been

[1]University of California, Santa Cruz, CA, USA
[2]Gargani + Company, Inc., Berkeley, CA, USA
[3]Bahçeşehir University, Istanbul, Turkey

Corresponding Author:
Michael Strong, University of California Santa Cruz, Center for Research on the Teaching Profession, Mailstop Merrill Faculty Services, 1156 High Street, Santa Cruz, CA 95064
Email: mastrong@ucsc.edu

developed for all purposes use a variety of procedures such as charts, rating scales, checklists, rubrics, and narrative descriptions. The most widely used technique has been systematic classroom observation based on interactive coding systems. These allow the observer to record most of what the teachers and students do during a given time interval (Stallings & Mohlman, 1988). The coding systems strive to be objective and typically require few inferences or judgments on the part of the observer. Critics of this paradigm argue that it lacks a theoretical and conceptual framework and focuses merely on categories or behaviors that are easily observed with measurement instruments (Ornstein, 1995a, 1995b). On the positive side, the findings have provided a set of indicators of quality instruction that are claimed (but not proven) to be related to effectiveness as measured by student academic achievement (Brophy & Good, 1986; Rosenshine, 1987). Some of these aspects of classroom instruction are conducting daily reviews, presenting new material, conducting guided practice, providing feedback and correctives, conducting independent practice, and conducting weekly and monthly reviews.

The common methodological concerns about observational measures relate to reliability and validity, size or number of the teaching samples needed, how the data are analyzed, and generalizability across grade level or subject matter. An additional source of doubt is the shortage (and some may claim lack) of rigorous evidence that teachers who score high on a particular observational, normatively derived measure of effectiveness show equivalent success with regard to actual student learning. Perhaps the strongest evidence we have comes from multilevel studies in different settings that investigated whether teachers with high evaluation scores on the Framework for Teaching (Danielson, 2007), one of the most widely used measures of teacher quality that includes a classroom observation component, also have classes with correspondingly high student achievement gains (e.g., Gallagher, 2004; Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004). All of these studies documented that teachers with higher evaluation scores produced *slightly* larger learning gains in student achievement as measured by standardized tests. The most recent attempt to relate a modified form of Danielson's Framework to student achievement was in a study by Kane, Taylor, Tyler, and Wooten (2010) using observations in the Cincinnati schools. They found that an overall move from one basic level to the next was associated with about one-sixth of a standard deviation of student achievement gain, the strongest association thus far in such studies. It must be stressed, however, that these, at best, weak associations are based on more than classroom observations; to arrive at evaluations of particular teachers, a full assessment must be made across the four domains described by Danielson, which, in addition to frequent observations of classroom practice, depend on artifacts such as lesson plans and samples of student work. Consequently, the evidence thus far suggests that observations *may* be one of several inputs that contribute to a rich, intensive, and slightly predictive measure of teacher performance, but this relationship has not been demonstrated for observations alone.

Teacher Observation and Evaluation

Teacher observations often play a central role in formal teacher evaluations, usually along with other sources of information. Yet in spite of many years of effort by researchers to construct observational instruments for evaluating teaching, they are not thought of very highly as measures of accountability, for the most part because they are perceived as having low validity and are considered too cumbersome for routine use by busy principals. In fact, teacher evaluations in most settings incorporate only brief classroom observations and are used in unsystematic ways. As Peterson (2000) put it,

> At the same time that its development has been neglected, teacher evaluation is a widespread activity in the schools. In this activity, where good practice should be common, inadequate efforts and materials are the order of the day. Poor practice in teacher evaluation is quietly accepted, according to teachers, administrators, and researchers. Evaluations look about the same in district after district, and for teacher after teacher. When there are problems with bad teachers or bad evaluations, people talk about it like few other educational problems; the rest of the time teacher evaluation is ignored or disparaged. (p. ix)

There is not much evidence to suggest a strong relationship between these observation-based teacher evaluation ratings and student academic outcomes. Many studies across the years report quite small correlations between principal evaluations and student achievement. Medley and Coker (1987) summarized the research up to that point by stating,

> To this day, almost all educational personnel decisions are based on judgments which, according to the research, are only slightly more accurate than they would be if they were based on pure chance. (p. 243)

In their own study they focused on the accuracy of principals' judgments of teacher performance on three broad roles: imparting knowledge, encouraging good citizenship, and being a good colleague. They, too, found that the correlation between principals' performance ratings and learning gains was a poor .10 to .23. However, in this case principals' judgments were likely based on more than observations (because they knew the teachers); thus, the contribution, if any, that observations provided is unknown.

Jacob and Lefgren criticized these and some more recent research studies (e.g., Peterson, 1987, 2000) for being generally

based on small, nonrepresentative samples; not accounting adequately for measurement error; and relying on "objective measures of teacher performance that are likely biased" (Jacob & Lefgren, 2008, p. 104). In their own work, they conducted an analysis that compared principal evaluations of teacher effectiveness (performed specifically for their experiment) against levels of teacher education and experience as well as effectiveness based on student achievement gains. They found that principals could correctly identify teachers at the extremes of effectiveness but could not discriminate among those in the middle range. An interesting contribution of their study is their deliberation on the sources of information that principals use in making judgments about teacher effectiveness. These include, in addition to formal and informal classroom observations, reports from parents and student achievement scores. They point out that principals will differ in the degree of their sophistication in accessing data and in their interpretation of any signals they receive, and that these differences may be reflected in their ultimate judgments of the teachers' effectiveness. As in the Medley and Coker (1987) study, the principals would have based their judgments on more than observations, so again we cannot know the extent to which observations aided the result.

Cognitive Operations

One possible contributing explanation for the weak correlations between existing teacher observation instruments and teacher effectiveness as measured by student achievement is that their developers have not taken into account findings from psychology and cognitive science regarding the cognitive operations that influence judgments of human behavior. Researchers from these disciplines have identified phenomena such as *confirmation bias* (e.g., Wason, 1960), *motivated reasoning* (Kunda, 1990), and *inattentional blindness* (Mack & Rock, 1998; Simons & Chabris, 1999), all of which influence the way we observe. The first term describes a tendency to seek, embellish, and emphasize experiences that support rather than challenge already held beliefs, and the second suggests we look more skeptically at data that do not fit our beliefs than those that do. The third is a striking occurrence in which people fail to notice stimuli appearing in front of their eyes when they are preoccupied with an attentionally demanding task, as demonstrated in an experiment in which observers fail to notice a gorilla walking in front of a group of basketball players when they are focused on counting how many times a basketball is passed. These and other phenomena related to perception may adversely influence how we make judgments while observing a teacher's classroom (for a complete discussion, see Strong, 2009).

One useful way of conceptualizing these various cognitive operations is to frame them as two generic modes of cognitive function that describe what we might think of as intuitive versus deliberate or rational thought processes. Philosophers dating back to Socrates and psychologists over the past century (e.g., James, 1890/1950; Johnson-Laird, 1983; Neisser, 1963; Piaget, 1926; Vygotsky, 1934/1987) have conceptualized them in this manner. More recently, researchers have further emphasized and defined the distinction between these dual systems of cognitive processes: those executed quickly with little conscious deliberation and those that are slower and more reflective (e.g., Chaiken & Trope, 1999; Epstein, 1994; Kahneman & Frederick, 2002; Sloman, 1996). Stanovich and West (2000) called these "System 1" and "System 2" processes, respectively. The operations of System 1 are fast, automatic, effortless, associative, and difficult to control or modify, whereas those of System 2 are slower, serial, effortful, and deliberately controlled; they are also relatively flexible and potentially rule governed.

System 1 operations produce shortcuts, or heuristics, that allow us to function rapidly and effectively. A program of research studies (known now as the "heuristics and biases approach") conducted by Kahneman and colleagues has documented the persistence of systematic errors in the intuitions of experts, implying that their intuitive judgments may be endorsed, at least passively, by their rational processes from System 2, one of whose functions is to monitor the quality both of mental operations and overt behavior (Kahneman & Frederick, 2005). These studies suggest that the monitoring is normally quite lax and allows many erroneous intuitive judgments to be expressed along with the correct ones. Frederick (2005) demonstrated this clearly in an experiment in which he found that quite large percentages of highly intelligent college students failed to reject plausible but erroneous solutions to simple puzzles. The surprisingly high rate of errors in these easy problems illustrates how lightly the output of System 1 is monitored by System 2.

System 2 judgments are less often erroneous than System 1 judgments, and since the path to the result is conscious, errors can be corrected. Much of the unreliability in human judgment comes from our inability or disinclination to use System 2. This work is nicely described and summarized in Kahneman's (2002) Nobel Prize lecture. Thus, in a classroom setting we can imagine that an observer will generate both System 1 and System 2 judgments. We do not know how systematic and widespread is the influence of operations from System 1, but we may hypothesize that they could contribute to the apparent lack of success in predicting learning outcomes from observations of teaching behavior.

The Need for a New Teacher Evaluation Measure

The utility of being able to identify effective teachers is self-evident; the method for doing so is not. If, in spite of the lack of evidence, effective teaching can easily be observed and identified, then establishing this fact would be of great benefit to educators and allow us confidently to continue observing

teachers as we have. If it cannot, then we must ask how, if at all, we can improve the result. To this end, we report on three sequential experiments. Each established conditions that were progressively more conducive to making accurate classifications of teachers of known, disparate effectiveness. Our intention was to establish the limits of accuracy of judgments made by experts and nonexperts and, if that should prove lacking, to consider how we might develop a new observational measure that can accurately assess teacher effectiveness.

Experiment 1

Design

Using student achievement data from a school district's database, we calculated value-added scores for teachers.[1] We then identified teachers whose scores indicated that their effectiveness was consistently higher or lower than average over the prior three years. This resulted in two groups with a difference in mean value-added scores of roughly a 0.50 standard deviation. We randomly selected teachers from the high and low effectiveness groups, obtained their approval to participate, and filmed them during a regular lesson. We showed short excerpts of these films to judges from various backgrounds and asked them to decide to which group each teacher belonged and to explain the rationale for their choices.

This experimental design relies on two theoretical techniques, one from psychology and one from education. The first, the psychological technique of "thin slicing," has been adopted by an increasing number of researchers from diverse fields, and it derives from the consistent finding that judgments about other people made from short samples of their behavior, sometimes as short as a few seconds, tend to be highly predictive of judgments based on much longer samples. A thin slice is defined as a brief (i.e., shorter than five minutes) excerpt of expressive behavior sampled from the behavioral stream (Ambady & Gray, 2002). Previous work has demonstrated that thin slices can provide information about a range of psychological constructs, including dispositional characteristics, social relations, and job performance (for a review, see Ambady, Bernieri, & Richeson, 2000). This technique has been used in a variety of settings, including education (Ambady & Rosenthal, 1993), where researchers found that students could predict a professor's end-of-semester class evaluations from exposure to a few seconds of his or her filmed lecturing. More recently, Benjamin and Shapiro (2009) used thin slicing to test naïve subjects' ability to forecast the outcomes of gubernatorial elections by viewing short clips of their debates, finding them to be more accurate in their predictions than models based on economic circumstances.

This body of research demonstrates that it is possible to obtain dependable ratings from a large number of participants without requiring lengthy laboratory sessions. In the present study, it meant that we could show observers short segments from filmed lessons and still be confident that the resulting judgments would be highly indicative of those based on viewing films of the whole lessons. Consequently, the penalty that we should expect to pay for using thin slices, if any, is in terms of the precision of judgments, not their overall accuracy. It is important to note that thin slicing is not far removed from what often occurs and what is sometimes advocated in the real world when school principals evaluate teachers. See, for example, *The Three-Minute Classroom Walkthrough* (Downey, Steffy, English, Frase, & Poston, 2004). However, the appropriateness of thin slicing for this experiment is not intended to suggest an endorsement of short walkthroughs as a viable or reliable method for administrators to evaluate teachers.

The second technique, value-added modeling (VAM), has been used by educational researchers, policy makers, and administrators to estimate the effects of teachers or schools on the learning of students. In this context, learning is almost always measured by gains on standardized achievement tests. Researchers such as William Sanders and his colleagues (e.g., Wright et al., 1997) maintain that VAM demonstrates the importance of teachers as a source of variance in student learning outcomes, and VAM is considered by many, including the U.S. Department of Education, to be a promising method to estimate teachers' effectiveness as defined by their contributions to student achievement gains. Acknowledging the limitations of VAM and the controversies surrounding it, we employ VAM scores in this study with the justification that they are widely used, that the standardized tests on which they are based are highly relevant measures for policy makers in the United States, and perhaps most importantly that our purpose is to estimate a global relationship between VAM scores and the judgments of observers. Consequently, our conclusions are not based on the ability of individual observers to make highly accurate and reliable classifications of individual teachers; rather, they depend on whether observers as a group tend to do better than chance when classifying a number of teachers. Because of this, our conclusions are less affected by any bias or lack of precision that may be inherent in our VAM scores than the conclusions about individual teachers that have been at the center of the VAM controversy (see Note 3). Our purpose is further aided by the fact that we created an experimental contrast between the high and low effectiveness groups that should be easily noticed. A difference of 0.50 standard deviations, the minimum magnitude of the difference in mean VAM scores between the groups, has been characterized as one that is large enough to be visible to the naked eye (Cohen, 1988, p. 26). Because the causal relationship between instruction and achievement is a noisy one, the corresponding difference in instruction should be even larger and thus more noticeable.

Selection of Teachers

The teachers whom we filmed worked in a medium-sized California school district that conducted annual testing and maintained a database of student test scores linked to teachers by unique identifiers, enabling us to estimate value-added scores. Fourth-grade teachers who had a three-year history of classes that performed at least one-half a standard deviation above the mean in math value-added gains constituted the high group. The low group consisted of fourth-grade teachers whose classes had not achieved gains of at least one-half a standard deviation above the mean in any of the previous three years. This identification produced a possible pool of almost 30 teachers, 10 of whom were randomly selected for participation, allowing for some with higher and some with lower performance records. If a teacher declined the offer, he or she was replaced with a randomly drawn substitute.

Researchers contacted the teachers and offered them $100 as compensation for permitting us to film a lesson on fractions. Ten White, female teachers formed the final sample of those who agreed to be filmed. One subsequently withdrew her consent, and two films were of poor quality, leaving us with seven teachers in the sample, three from the high and four from the low group, who were separated by 0.50 standard deviations or more in the VAM scores. The seven teachers had between seven and 19 years of classroom experience. Age and experience were equally distributed across high and low groups. All had elementary teaching credentials, and none were math specialists.

Selection of Judges

A total of 100 judges took part in Experiment 1. These were distributed among 10 categories: school administrators, education professors, math educators, teacher educators, parents of elementary school children, K-12 teachers, undergraduate students taking education courses, teacher mentors, elementary school children, and adults with no formal connection to education. The purpose of selecting judges from different backgrounds was to determine if their relationship to education affected their judgments. Judges were recruited through various means including personal invitation, flyers distributed in educational establishments, and word of mouth. Recruitment continued until we had administered 100 sessions distributed across the 10 categories of judge.

Experimental Materials

We extracted two-minute segments from each of the seven films using a commercial computer digital editing program. At this point all researchers except the statistician who calculated the value-added scores were still blind to the group affiliation of the teachers. Each clip featured the teacher presenting part of a lesson on fractions to the whole class. In most cases, we selected the first two consecutive minutes of whole class instruction that occurred in the lesson.[2] Exceptions were made if an interruption occurred such as the phone ringing or a behavior management episode, in which case the next two consecutive minutes were chosen. In all cases these clips occurred no sooner than 11 minutes into the 50-minute lesson and no later than 19 minutes before the end of the lesson. Clips were labeled with a teacher number. We constructed 10 playlists using different random orders of the seven clips. We designed score sheets listing each teacher by number with a small still photo and an option for the observer to check "yes" if that teacher were judged to be in the high group or "no" for the low group. The exact instructions are available in Appendix A. A debriefing interview (see Appendix B) questioned the judges about their educational background and experience, the criteria they used to make their selections, their confidence in their judgments, and their perceptions about the task itself.

Experimental Procedures

A researcher, blind to the group affiliation of the teachers, administered the experiment individually with each of the 100 participants in any available quiet setting, usually a private office, reading the exact instructions from Appendix A. Each participant then viewed a randomly selected playlist of the seven clips on a notebook computer using a headset. Each clip was separated by a 30-second pause for completing the score sheet, where participants recorded whether the teacher was in the above-average group and any notes they wished to help them in their judgments. They were told that the clips had been randomly chosen from a larger sample and that any clip could be of a teacher in either group. After the participant had viewed all seven clips and completed the score sheet, the researcher conducted the debriefing interview lasting about 15 minutes. This interview was audio-recorded.

Research Questions

1. *Agreement.* Among observers from a variety of backgrounds, what is the interjudge agreement when evaluating teachers through observation of thin slices of teaching behavior? Does agreement vary by teacher or category of judges?
2. *Accuracy.* Among observers from a variety of backgrounds, what is the level of judges' accuracy when evaluating teachers of known levels of effectiveness through observation of thin slices of teaching behavior? Are experts more accurate than nonexperts? (Throughout this article, by "accuracy" we mean the ability to distinguish teachers with above-average VAM scores from those who were average or lower.)

3. *Criteria.* What criteria do judges from a variety of backgrounds use when evaluating teachers through observation of thin slices of teaching behavior?
4. *Confidence.* How confident are judges that they made accurate assessments of teacher effectiveness?

Results

Agreement. Several questions related to interjudge agreement may be posed. First, are individual judges likely to agree with one another? We answered this by calculating the intraclass correlation (ICC) that Shrout and Fleiss (1979) refer to as ICC(2,1), which estimates the interjudge reliability of the scores of individual judges. This was low (.24), which is not unexpected given the observers were not trained and were free to use their own criteria for making judgments. However, this estimate is greater than zero, which indicates that scores are systematic, albeit shrouded in considerable noise.

This raises a second and, for our purposes, more relevant question: How well do ratings agree when we aggregate them across judges? We answered this by calculating ICC(2,100), which estimates the reliability of a mean score computed from all available individual scores. This was high (.97), indicating that we have a sound basis for estimating a global relationship between VAM scores and the judgments of observers as a group. The validity of estimating this global relationship is also demonstrated by the median pairwise Pearson correlation of (a) scores provided by individual raters and (b) the average scores for the 10 categories of judges. These values are .167 and .835, respectively, indicating again that as a group judges were quite reliable.

A third question is whether the level of agreement varied. Table 1, which indicates that the rate of agreement by teacher ranged from 63% to 84%, suggests that agreement varied across those being judged. Likewise, Table 2, which indicates that the agreement across categories of judges ranged from 62% to 83%, suggests that agreement varied across groups providing the judgments. Explaining and minimizing this variation is an important part of our effort to develop a more highly predictive observational measure.

Accuracy. Again, we can ask several questions about the accuracy of judges. First, are judges accurate overall? A simple count of the correct assignments by each judge, presented as a histogram in Figure 1, suggests that they are not. Possible scores for any one judge ranged from zero (none correct) to seven (all correct). The mean number correct was 2.8 and the mode 3.0, both of which are lower than chance (3.5).

To investigate this question further, we fit a three-level hierarchical generalized linear model (HGLM) with a logit link (see Raudenbush & Bryk, 2002, Chapter 10) in which judgments about teacher effectiveness were nested within judges who were nested within categories of judges. The model took the form

Table 1. Experiment 1: Interjudge Agreement (N = 100) by Teacher

Teacher #	Group	% agreement
1	H	65
2	H	78
3	L	63
4	L	68
5	H	84
6	L	67
7	L	80

H = high effectiveness group; L = low effectiveness group.

Table 2. Experiment 1: Judges' Agreement by Group

Group	*n*	% agree
Teachers	10	83
Parents	7	80
Mentors	10	79
University professors	9	78
Administrators	10	77
Teacher educators	10	77
College students	11	75
Math educators	10	74
Other adults	11	70
Elementary students	12	62

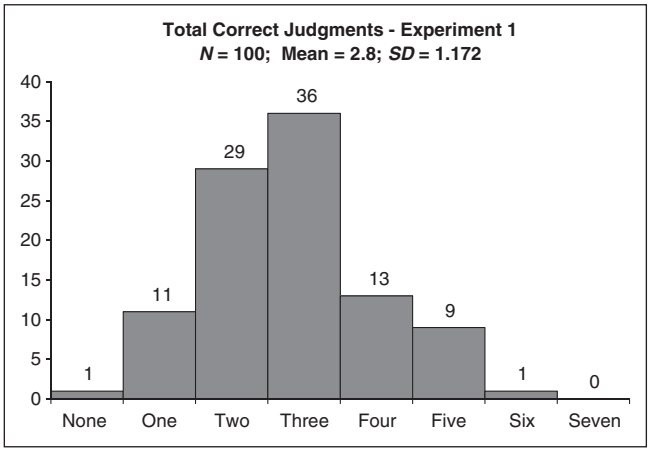

Figure 1. Distribution of total correct judgments: Experiment 1

$$\eta_{ijk} = \gamma_0 + \sum_{t=1}^{(T-1)} \left(\gamma_t X_{tijk}\right) + u_{jk} + r_k, \quad (1)$$

where η_{ijk} is the log of the odds (logit) that judge j from category k correctly placed teacher i in either the high or low effectiveness groups, γ_0 is the overall accuracy of judges expressed as the log of the odds of making a correct judgment, the X_{tijk} are effect-coded variables indicating to which of the T teachers the score was assigned (with one teacher acting as

a reference category), γ_t provides information on the differential accuracy of judges across teachers, u_{jk} is a random error term associated with judges that is distributed $N(0, \sigma_2^2)$, and r_k is a random error term associated with the categories of judges that is distributed $N(0, \sigma_3^2)$.

If judges were guessing, we would expect our estimate of their overall accuracy, γ_0, to reflect chance and thus be close to 0 logits (equivalent to 50% accuracy). Using the model above, we estimated γ_0 to be –0.486 logits (38% accuracy), which is statistically significantly lower than chance (SE = 0.088, t = –5.502, df = 99, $p < .0005$). Thus, the best strategy for identifying effective teachers in this case would be to place them in the *opposite* categories to which judges assigned them, resulting in correct classifications almost two-thirds of the time.

Another important question is whether judges were equally accurate (or inaccurate) when judging each teacher. Again, a simple count brings this into doubt. We estimated the percentage of correct responses for all judges by teacher and for each group overall (Tables 3 and 4). Three of the teachers (one high and two low) were accurately rated by between 63% and 67% of the judges, but four were accurately rated by only 16% to 32% of the judges. Curiously the highest agreements from Table 1 were for teachers who were inaccurately assigned to the high or low group.

We used the estimates of γ_t from the HGLM model above to establish more rigorously differential accuracy across teachers. Table 5 presents these estimates as probabilities, odds ratios, and logits. For six of these values, t tests were performed to determine if the accuracy for individual teachers deviated from the overall accuracy across all teachers (one estimate was constrained by the model to be a function of the other estimates, so statistical inference is not possible). Because we were conducting simultaneous inference, we corrected the cutoff value for statistical significance using a Bonferroni adjustment, which in this case was .05/6 = .0083. In five of the six cases, the accuracy of judges with respect to individual teachers was statistically significantly higher or lower than their overall accuracy, providing more evidence that the judgments were systematic rather than random.

A third question is whether judges in some categories were more accurate than others. We can see from Table 4 that the elementary students reached chance levels of correctness (probably because they really were guessing), while all other groups performed below chance, suggesting that their educational knowledge and experience interfered in some way with their abilities to identify the more effective teachers. Administrators, teacher educators, and math educators were accurate only about one-third of the time. This contradicts the hypothesis that experts would outperform the non-experts. To investigate this further, we expanded the HGLM model above to determine if judges with educational expertise (administrators, mentors, teacher educators, professors of education, and math professors) were more accurate. We added a dummy-coded variable (*Expert*) to the model

Table 3. Experiment 1: Interjudge Accuracy (N = 100) by Teacher

Teacher #	Group	% correct
1	H	65
2	H	22
3	L	63
4	L	32
5	H	16
6	L	67
7	L	20

H = high effectiveness group; L = low effectiveness group.

Table 4. Experiment 1: Judges' Accuracy by Group

Group	*n*	% correct
Elementary students	12	50
Mentors	10	47
Other adults	11	43
University professors	9	41
Teachers	10	37
Parents	7	37
College students	11	36
Math educators	10	34
Teacher educators	10	31
Administrators	10	31

that indicated whether the group to which the judge belonged was one that had educational expertise, such that

$$\eta_{ijk} = \gamma_0 + \sum_{t=1}^{(T-1)} \left(\gamma_t X_{tijk}\right) + \gamma_T (Expert)_k + u_{jk} + r_k . \quad (2)$$

The resulting γ_T term is an estimate of the difference in accuracy of educational experts and nonexperts, and its value (–0.166 logits) is close to zero and not statistically significant (SE = 0.178, t = –0.933, df = 8, p = .378). This provides evidence that whatever was driving the systematic nature of the judgments and their inaccuracy, it was something to which educational experts were not immune.

Judgment criteria. The factors influencing subjects' judgments fell into four distinct categories: student engagement, teaching strategies, teacher characteristics, and math knowledge. The most frequently cited indicator was the level of student engagement. This was mentioned by all the respondents in the teacher, administrator, mentor, and parent categories; more than 80% of the other adults, math educators, and university professors; more than 70% of the teacher educators; and more than half the college students. (The elementary students were omitted from this part of the analysis since they produced no reasons for their selection beyond "Because I liked her" or "I don't know.")

Table 5. Estimates from the Hierarchical Generalized Linear Model (HGLM) in Equation 1

		Coefficient Estimates and Hypothesis Tests						
	Parameter	Estimate (Logits)	*SE*	*t* value	*df*	*p* value	*OR*	Proportion accurate
Overall accuracy	γ_0	–0.486	0.088	–5.502	9	.000	0.615	.381
Teacher 1	γ_1	–0.780	0.222	–3.514	693	.001	0.458	.220
Teacher 2	γ_2	0.975	0.195	5.002	693	.000	2.651	.620
Teacher 3	γ_3	–0.362	0.204	–1.772	693	.076	0.696	.300
Teacher 4	γ_4	–1.100	0.241	–4.557	693	.000	0.333	.170
Teacher 5	γ_5	1.104	0.198	5.588	693	.000	3.018	.650
Teacher 6	γ_6	–1.031	0.237	–4.354	693	.000	0.357	.180
Teacher 7	(Constrained)	1.194	—	—	—	—	3.299	.670

Table 6. Experiment 1: Teaching Strategies Commonly Cited as Influencing Judgments

Strategy	%
Accesses students' prior knowledge	67
Has active interaction with students	36
Moves around classroom	34
Enables students to generate ideas	23
Creates stimulating classroom environment	23
Uses visuals and manipulatives	23
Checks for student understanding	18
Has clear objectives	14
Presents concepts clearly	11
Exhibits equity	8
Differentiates instruction	7

Table 7. Experiment 1: Responses to Questions Related to Judges' Confidence

Group	% same judgment in one minute	% not surprised if correct
Math educators	78	44
Other adults	67	22
Mentors	56	44
College students	55	0
Teachers	22	33
Parents	20	40
Teacher educators	14	14
Administrators	13	25
University professors	0	33
Mean	40	27

Eleven different teaching strategies were mentioned with some frequency in the debriefing sessions, and these are listed in Table 6. Fully two-thirds of the participants commented that one or more teachers accessed students' prior knowledge (often by relating the concept of fractions to the lives of the students), or failed to. The next strategy was mentioned only just more than half as frequently.

The subjects referred to two kinds of teacher characteristics. Forty-five percent mentioned the teachers' confidence, energy, or "presence," and 10% valued a teacher with a sense of humor or an engaging personality. These characteristics to do with outgoingness or expressiveness have been found to be associated with positive evaluations of college professors (e.g., Naftulin, Ware, & Donnelly, 1973; Radmacher & Martin, 2001; Ware & Williams, 1975; Williams & Ceci, 1997).

Confidence. Two debriefing questions were related to participants' confidence with the experiment. One addressed the issue of the brief exposure to the teaching behavior, asking participants if they felt they would have been able to make the same judgment had they watched for one minute instead of two. The second asked if they would be surprised if they found out that their judgments were 100% correct. Responses to both questions varied widely, with an average 40% of the subjects in the main sample feeling one minute would have been sufficient, and 27% reporting they would not be surprised if they were completely correct (Table 7). Degree of confidence appeared to have no relation to degree of accuracy.

Discussion

The impressive rate of agreement among judges overall suggests that regardless of background, judges were responding to systematic influences. At the same time, the judgments made by both expert and nonexpert judges were inaccurate in ways that also reflected systematic influences—certain teachers were inaccurately rated by a significant majority of judges while others were not, and the accuracy of judges overall was significantly lower than would have been produced by chance. Given this, it appears that nonrandom influences, possibly System 1 operations, appear to have led judges astray. There are, of course, other possible explanations for these findings; we enumerate the most salient alternatives below and strive to rule them out in the experiments that follow.

1. *Biased VAM scores.* There is reason to believe that VAM scores estimated with linear regression, such as

the ones we used, can be biased (McCaffrey, Han, & Lockwood, 2009). This might cause accurate judges to appear erroneously to be inaccurate. Although we have reason to believe that this is not a problem given our analysis of responses,[3] we cannot entirely rule out the possibility that the VAM scores may have affected our results.

2. *Too small a contrast between high and low performing teachers.* It is possible that the difference in VAM scores between the high and low performing groups of teachers may not have been large enough for judges to discriminate. All the teachers in the high performing group had VAM scores that were at least 0.5 standard deviations above the district mean for three years. On the other hand, teachers in the low performing group had VAM scores that were consistently below the district mean, not 0.5 standard deviations below the mean. Thus, it is possible that judges were being asked to distinguish between pretty good teachers and very good teachers, which may have proved too subtle a distinction and one that did not correspond well to the labels "high" and "low" performing.
3. *Inadequately trained judges.* Judges may have lacked the training—on a specific observation protocol or more broadly in education—that they needed to be accurate. This might have led judges to guess, apply criteria inconsistently, or rely on irrelevant criteria. These potential problems could have been heightened by selecting relatively small numbers of judges from heterogeneous groups. The relatively high agreement across raters, however, argues against guessing or the inconsistent use of criteria, and performing worse than chance suggests that their criteria were not only irrelevant but misleading. Nonetheless, a larger and better trained group of judges might provide more accurate assessments.
4. *Nonrepresentative video clips.* The video clips may not have represented the true instructional style of teachers for two reasons. First, given the small sample of teachers, it is possible that the randomization procedures, by chance, failed to produce a set of clips that were representative, thereby leading judges to appear more inaccurate than they were. The fact that teachers were selected at random from a pool of potential subjects, that the researcher selecting the clips did not know the group affiliations of the teachers, that the clips were chosen according to the same procedures for each teacher, and that the clips were presented in random order protects against systematic bias, but it does not rule out the possibility that a particular sample is biased. Second, although we can be confident from previous research that judgments made from short exposure to teaching behaviors are likely to correlate highly with judgments made from longer observations, it is possible that the clips were too short to be representative.
5. *Changes in the student population.* It is conceivable that the classes we filmed (comprised of students from the year of the study) were systematically different from the classes used to estimate value-added scores (comprised of students from the three years prior to the study). This in turn could have led teachers to adopt atypical styles of instruction or to be more or less effective with their new classes than they had been historically. Although there is no reason to believe that this was the case, it is possible and it had the potential to make judges appear less accurate than they really were.
6. *Idiosyncratic local context.* For any number of reasons not described above, it is possible that our results accurately reflect the reality of the school district in which the experiment was conducted but that the school district is so unusual that the results do not apply elsewhere. We have no reason to believe this is the case, but as with all experiments replication is essential.

Experiment 2

Design and Methodology

To rule out some of the alternative explanations of Experiment 1, we replicated the experiment—with a number of intentional differences—using new samples of teachers, film clips, and judges. First, we selected teachers from a different district located in a different state: Tennessee (to help rule out Alternative Explanation 6). Second, we selected teachers based on a three-year record of performing at least 0.50 standard deviations above *or below* the district mean value-added score (to help rule out Alternative Explanation 2) and calculated the scores in a way that was less prone to bias[4] (to help rule out Alternative Explanation 1). Then we filmed 20 fourth- and fifth-grade teachers giving a lesson on fractions and selected four above average and four below average for the experiment, based on the quality of the films and the similarity of the curriculum covered in the lesson (to help rule out Alternative Explanation 4). There were six female and two male teachers, most in the middle of their careers, and one approaching retirement. To increase the chance of finding a relationship, we showed the clips to more judges (165) who all had expertise in education—school principals, assistant principals, and administrators-in-training—drawn both from Tennessee and California (to help rule out Alternative Explanation 3). The films were presented either in group format or from a secure website via the Internet, using a proprietary software program requiring a unique password good for a single use. To confirm the accuracy of the group affiliations of the teachers, we checked their classes' achievements

for the year of filming and found them to be consistent, in all eight cases, with their original standings derived from the historical data (to rule out Alternative Explanation 5).

Research Questions

1. *Agreement.* Among school administrators or administrators-in-training, what is the interjudge agreement when evaluating teachers through observation of thin slices of teaching behavior?
2. *Accuracy.* Among school administrators or administrators-in-training, what is the level of judges' accuracy when evaluating teachers of known levels of effectiveness through observation of thin slices of teaching behavior? Are those with more expertise more accurate?

Results

Agreement. The patterns of interjudge agreement among the administrators were similar to those found for judges in the first experiment. The ICC(2,1) and ICC(2,165) estimates were .27 and .98, respectively, which are very close to those from the Experiment 1. As before, agreement across teachers varied considerably, ranging from 60% to 88% (Table 8).

Accuracy. As in Experiment 1, a simple count of correct assignments, presented as the histogram in Figure 2, reveals that judges were not accurate; judges classified from zero to seven teachers correctly, and the mean number correct, 3.85 out of 8, was slightly less than would be produced by chance. As before, we investigated this further by fitting an HGLM, but because all judges were school administrators we used a two-level version of the model described above, such that

$$\eta_{ij} = \gamma_0 + \sum_{t=1}^{(T-1)} \left(\gamma_t X_{tij}\right) + u_j . \quad (3)$$

In this case, the overall accuracy of the judges, γ_0, was estimated to be slightly less than chance at –0.102 logits (47% accuracy), but this underperformance was not statistically significant ($SE = 0.074500$, $t = -1.369$, $df = 164$, $p = .173$).

Accuracy again varied by teacher; two teachers, one high performing and one low performing, were accurately rated by more than 80% of the administrators, but two other teachers were accurately rated by only 12% and 21% of the judges (see Table 9). Estimates of γ_t bear out this differential accuracy: Seven out of seven estimates were statistically significantly different from the overall rate of accuracy after making a Bonferroni correction (see Table 10).

We also investigated whether administrators with more expertise (those who completed their training and worked in the field) were more accurate than those with less expertise (administrators who were in training). We expanded the model given in Equation 3 to include a dummy variable for expertise, such that

$$\eta_{ij} = \gamma_0 + \sum_{t=1}^{(T-1)} \left(\gamma_t X_{tij}\right) + \gamma_T (Expert)_j + u_j . \quad (4)$$

Table 8. Experiment 2: Interjudge Agreement ($N = 165$) by Teacher

Teacher #	Group	% agreement
1	H	67
10	L	60
14	H	81
3	H	62
5	H	79
12	L	82
4	L	65
7	L	88

H = high effectiveness group; L = low effectiveness group.

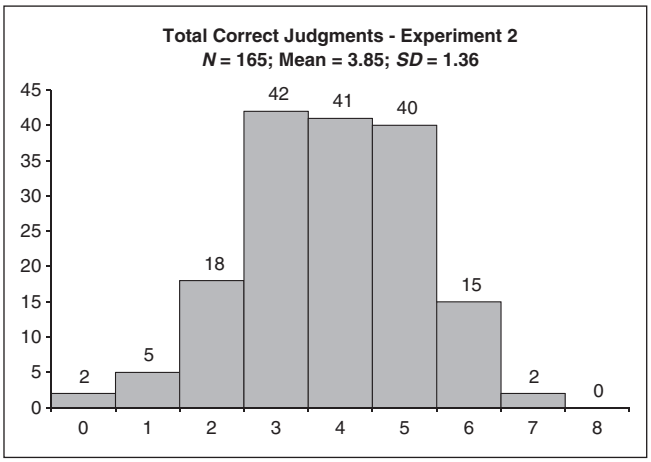

Figure 2. Distribution of Total Correct Judgments: Experiment 2

Table 9. Experiment 2: Interjudge Accuracy ($N = 165$) by Teacher

Teacher #	Group	% correct
1	H	33
10	L	60
14	H	81
3	H	62
5	H	21
12	L	82
4	L	35
7	L	12

H = high effectiveness group; L = low effectiveness group.

Table 10. Estimates from the Hierarchical Generalized Linear Model in Equation 3

		Coefficient Estimates and Hypothesis Tests						
	Parameter	Estimate (Logits)	*SE*	*t* value	*df*	*p* value	*OR*	Proportion accurate
Overall accuracy	γ_0	–0.102	0.075	–1.369	164	.173	0.903	.475
Teacher 1	γ_1	0.537	0.154	3.495	1312	.001	1.711	.607
Teacher 2	γ_2	1.355	0.175	7.730	1312	.000	3.876	.778
Teacher 3	γ_3	0.694	0.156	4.445	1312	.000	2.001	.644
Teacher 4	γ_4	–1.259	0.180	–7.001	1312	.000	0.284	.204
Teacher 5	γ_5	1.661	0.190	8.759	1312	.000	5.266	.826
Teacher 6	γ_6	–0.490	0.156	–3.140	1312	.002	0.613	.356
Teacher 7	γ_7	–1.954	0.222	–8.808	1312	.000	0.142	.113
Teacher 8	(Constrained)	–0.544	—	—	—	—	0.581	.344

The corresponding estimate of γ_T was −0.219 ($OR = 0.803$), which again was small and not statistically significant ($SE = 0.160$, $t = -1.369$, $df = 163$, $p = .173$). Interestingly, nonexperts were, as a group, close to chance (49% accuracy) and experts lower than chance (44% accuracy).

Discussion

The results of Experiment 2 essentially replicated those from Experiment 1. With a more experienced and larger group of judges, a different set of teachers grouped more widely apart according to their records of realizing student achievement gains, and a more robust set of value-added calculations, we still saw a large degree of overall interjudge agreement and accuracy that at best reflects chance.

These findings serve to strengthen the possibility that System 1 operations are overriding System 2 processes, even among school administrators who are likely to be experienced teacher evaluators. Furthermore, judges are responding to similar stimuli from the teaching behaviors, resulting in systematic (rather than random) evaluations that are not predictive of teacher effectiveness. However, certain teachers tend to be accurately identified by the majority of raters, suggesting that it might be possible to identify what it is about these teachers that aids judges and make use of that information in the development of a future measure. Other possible implications for the design of a measure that attempts to predict teacher effectiveness in student learning are that (a) users need to be trained to use System 2 rather than System 1 processes, (b) the measure should consist only of items that reliably distinguish more effective from less effective teachers, and (c) the measure should avoid items that trigger System 1 operations.

Experiment 3

Design and Methodology

As the next step in developing a more predictive measure, we replicated the experiment for a third time, but with three important differences—judges were well trained, used an established observational measure, and viewed the full-length videos of teachers (from Experiment 2) presenting their lessons—in order to rule out Alternative Explanations 3 and 4 more completely. We conducted this experiment in collaboration with researchers at the University of Virginia's Center for the Advanced Study of Teaching and Learning, to whom we submitted the eight full-lesson films for ratings using the CLASS instrument (Pianta, La Paro, & Hamre, 2008). CLASS consists of 11 dimensions across three domains that cover *emotional support* (classroom climate, teacher sensitivity, regard for student perspectives), *classroom organization* (behavior management, productivity, instructional learning formats), and *instructional support* (concept development, quality of feedback, language modeling) and a fourth dimension of student engagement. The CLASS instrument is organized with student engagement as the dependent variable and is believed to measure the effectiveness of teachers. (For technical information on the reliability and validity of CLASS, see the *CLASS Manual*; Pianta et al., 2008). Trained raters, ignorant to the value-added histories of the teachers (but aware that some were above and some below average), viewed and double-coded the eight lessons and scored them using the CLASS protocol. This produced a set of total scores that enabled a ranking of the teachers relative to each other. To assess accuracy, we allowed the scoring in the top half of the rankings to indicate above-average effectiveness, and in the bottom half to indicate below-average effectiveness. We also compared the CLASS ranks to those produced in Experiment 2, which were calculated according to the number of nominations a teacher received for being in the above-average group.

Research Questions

1. *Accuracy.* For well-trained judges using the CLASS protocol, what is the level of judges' accuracy when evaluating teachers of known levels of effectiveness through observation of teaching behavior during a full lesson?

Table 11. Rankings (1 = *Best*) of Eight Teachers in Experiment 2: Administrators and by CLASS

Teacher	Group	Rank two-minute judges	Rank CLASS
#14	H	2	2
#1	H	6	7
#3	H	4	4
#5	H	8	5
#10	L	5	3
#4	L	3	6
#12	L	7	8
#7	L	1	1
Total correct		4	4

H = high effectiveness group; L = low effectiveness group.

2. *Replication.* Do well-trained judges using the CLASS protocol and viewing a full lesson produce different results than school administrators using personal judgment after viewing two-minute clips of the same lessons?
3. *Discrimination.* Which if any items from the CLASS protocol effectively discriminate high and low performing teachers?

Results

Accuracy. As presented in Table 11, judges using the CLASS protocol correctly categorized 50% of the teachers. This result is indistinguishable from chance using any statistical test.

Replication. The Spearman rank correlation for the values in Table 11 is .714 and is statistically significantly different from zero ($t = 2.5$, $df = 6$, $p = .047$). Thus, even with additional training, a structured observational protocol, and access to the full lesson, judges replicated the results of Experiment 2.

Discrimination. Although the overall CLASS scores did not successfully discriminate between the teachers in the above and below average groups, we wondered if certain items in the measure might discriminate between these groups. To investigate this, the trained scorers rated the lessons of the 12 teachers from Tennessee who were not included in Experiment 2, still ignorant as to their group affiliation. Then, after learning of the teachers' groupings, the raters did an item-by-item comparison of all 20 teachers divided into their two groups. This analysis showed that a small subset of items produced scores that accurately identified teachers as either above or below average. All of these items were from the instructional domain. They included clearly expressing the lesson objective, integrating students' prior knowledge, using opportunities to go beyond the current lesson, using more than one delivery mechanism or modality, using multiple examples, giving feedback about process, and asking how and why questions.

Discussion

The results from the CLASS ratings raise new questions and add weight to some of the possible interpretations of the findings from the first two experiments. Foremost, why would there be little difference between the ratings from Experiment 2 and Experiment 3? It may be that CLASS was not designed to be predictive of a teacher's value-added gains, per se, but an alternative or more expansive conception of good teaching. Furthermore, this conception may reflect a more formalized arrangement of the criteria used by judges in Experiments 1 and 2. CLASS is focused primarily on student engagement, and nearly two-thirds of its items gauge emotional support and classroom organization; the most commonly cited rating criteria in Experiment 1 was student engagement, followed by teaching strategies, teacher characteristics, and math knowledge. On the face of it, these appear to reflect very similar concerns. Yet the CLASS items that were identified as predictive of teacher effectiveness came from the one-third that gauges instructional support. This suggests that emotional support and classroom organization may be highly valued and even necessary, but not sufficient to ensure effective teaching. An instrument that is designed to predict student achievement would therefore do better to focus solely on the instructional items, which should yield low scores in classes where positive organization and climate are lacking but not necessarily high scores when those features are well established.

Another possibility is that System 1 operations may influence judges even when they are trained to use a structured, rigorously developed instrument. Thus, it may not be enough just to have an instrument in hand, in contrast to the judges in Experiment 2 who had nothing; it may be necessary to train observers to be aware of and to factor out the many cognitive influences on judgments that result in errors, omissions, and inaccuracies.

Conclusion

This study reports on three experiments in which judges viewed teachers giving lessons and then categorized teachers as either high or low performing, defined in all cases by the achievement gains of their classes on standardized tests. Our purpose was to examine whether, under increasingly more favorable conditions, judges agreed with one another and were accurate. In every case, judges achieved relatively high levels of agreement but were absolutely inaccurate, leading us to question whether educators can identify effective teachers when they see them. This in turn has motivated us to undertake development of an observational measure that can predict teacher effectiveness. The experiments we described suggest that we can by focusing on some aspects of the instructional domain suggested by the CLASS protocol while taking System 1 operations into account.

An important dimension of this work concerns values—that is, how should we define effective teaching? We recognize that students and their parents hope that schools will provide a wide range of benefits. We have chosen to define educational effectiveness narrowly as the value a teacher adds to gains in student learning as measured by standardized test scores. We share some of the skepticism surrounding value-added scores, yet we acknowledge that they reflect something that policy makers and much of the public truly value, and we believe that at least in aggregate they measure what they claim to measure. Nonetheless, we also believe it is important to use other methods for measuring student learning. Regardless of how student learning is defined, in the present climate of Race to the Top and other accountability initiatives, the search for effective teachers has become widespread, urgent, and high stakes, and it is our hope that an observational measure that reliably identifies the best teachers will help administrators find and support them, and in so doing advance the cause of education in a meaningful way.

Appendix A

Instructions for Raters in Experiments 1 and 2

You will view seven 2-minute clips of elementary teachers giving a math lesson. There is no break between clips. Some of these teachers have classes who have above-average success in learning year after year. Others have classes with average success in learning. For each clip, please circle "yes" on the score sheet if you think the teacher was in the above-average group or "no" if you think the teacher was in the average group. Make sure you match your rating with the correct teacher number. You may change your mind at any point, if you wish.

Appendix B

Debrief Protocol for Experiment 1

THE RATINGS

1. Some of the teachers you judged to be successful. What was it about these teachers that made you judge them this way? [Provide screen shots of teachers.] Could you say something about them as a group, and then talk about the individual teachers?

   PROBE for further information or illustrative examples if they use any descriptors such as "charisma," "presence," "good questioning," "engagement," etc.
2. Here are the other teachers. [Provide screen shots.] Can you talk about what you saw in their clips that caused you to judge them as you did?

   Again PROBE where necessary.
3. In general, thinking about all the clips, what criteria did you find yourself using to rate the teachers?

   Prompt for the underlying meaning systems, values, purposes of education, and illustrative examples.
4. Why did you rely on these criteria?
5. If you could have had one other piece of information to make your decision, what would that be? Why?
6. What do you suppose is the value of using the criteria you have mentioned?
7. Are there any limitations of using these criteria? If so, what are they?

THE TASK

1. In what ways was the exercise easy or difficult? Please explain and give examples.
2. If you had watched only one minute for each teacher, could you have made the same determination? What about 30 seconds? Do you think you may have changed your judgment if you had watched the teachers for 10 minutes or a full lesson?

   PROBE for reasons why or why not.
3. What if I told you your judgments were 100% correct—would that surprise you? To what would you attribute that success?

   PROBE for deeper explanation, sources, or relevant experience.
4. What if I told you your ratings were mostly wrong—would that surprise you? To what would you attribute that lack of success?
5. What do you suppose this exercise tells you about teaching?
6. What do you suppose this exercise tells you about teacher quality?
7. Is there anything else this exercise makes you think about that you'd like to mention?

Acknowledgments

Thanks to *Journal of Teacher Education* reviewers for their thoughtful comments. Thanks to Bob Pianta, Jennifer Locasale-Crouch, Emily Davis, and Wendy Amato of the University of Virginia for their generous and important collaboration.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: This article is based on work funded by a grant from the Carnegie Corporation of New York to the first author.

Notes

1. Calculation of teacher value-added scores from the student achievement data was done using a regression equation that included student and class characteristics. The analysis was guided by Sanders' work (Sanders & Horn, 1995) on value-added models in that we used the previous year's student achievement test score to predict the current year's score. We chose to use a simple regression equation to analyze the data. As there were only three variations in new teacher support programs, we chose to treat each district as a case study of the new teacher support, focusing particularly on the variation of mentor-to-novice ratio.

   A conceptual description of the regression equation can be written as

   $$\textit{Current Score} = \textit{Constant} + \textit{Previous Score} + \textit{Student Factors} + \textit{Class Factors} + \textit{School/External Factors}.$$

   For this analysis, we defined student factors in terms of a student's minority and poverty status. Student minority status is defined as one if a student is an ethnic minority or zero if Caucasian. Student poverty level is one if a student participates in the free/reduced-cost lunch program and zero if not. Class factors include the proportion of students in a class who are of minority status and who are receiving free/reduced-cost lunch as well as the class's level of prior achievement. We also included a dummy variable for school in the equation as a way of separating out school- and district-level variance. We recognize that this method of calculating value-added scores has been criticized for yielding biased estimates. It is a limitation of this part of the study that we addressed in the second and third experiments.
2. We recognize that activity structures other than direct instruction exist and are often preferred. However, we chose to use examples of direct instruction for the experiment because it was easier to standardize across teachers and guaranteed that raters would see a teacher interacting with the whole class with a range of behaviors that included explaining concepts, fielding questions, and giving and eliciting examples.
3. The possibility that bias affected our results is not readily supported by the pattern of responses we observed. Bias would manifest itself as the regression line being systematically shifted up or down or the slope being systematically too flat or too steep. In the case of a shifted line, our VAM scores would erroneously lead us (the researchers) to believe that on average teachers were more (or less) effective than they really were. The proportion of teachers classified as highly effective by accurate judges would then be less than (or greater than) the proportion of teachers classified by the researchers as highly effective. This was not the case. Using the first hierarchical generalized linear model given in Equation 1, this time with the binary outcome variable equal to one if the teacher was judged highly effective and zero otherwise, the observed overall proportion of teachers classified as highly effective was estimated to be .47 ($\gamma_{000} = -0.127$ logits). This was very close to the actual proportion of highly effective teachers of .43 (–0.288 logits), and a hypothesis test of whether the observed less the actual differed from zero was not statistically significant (–0.127 + 0.288 = 0.161 logits, $SE = 0.088$, $t = 1.84$, $df = 9$, $p = .100$).

   In the case of a flattened regression line, we (the researchers) would underestimate the difference in VAM scores between high and low effectiveness groups of teachers. Thus, what we estimated to be a 0.50 standard deviation difference in VAM scores would in fact be a greater difference in effectiveness, and that would be associated with an even more easily observed difference in instructional styles. If this were taking place, we would expect the accuracy of judges to be greater than chance. On the other hand, if the regression line were systematically too steep, the difference between the two groups would diminish and judges would be reduced to guessing. If this were taking place, we would expect the accuracy of judges to be at about chance. However, the accuracy of judges was below chance, which does not easily reconcile itself with the consequences of using biased VAM estimates.
4. Value-added scores were provided to us by the school district, which, as do all districts in the state, receives its value-added scores from William Sanders and the SAS Institute in North Carolina. This statement from the SAS website describes Sanders's approach to calculating value-added scores:

   > To accommodate the technical requirements of a mixed-model application of the scope of SAS EVAAS, Sanders and his colleagues have developed a software system capable of solving thousands of equations iteratively. This complex system enables a massive multivariate, longitudinal analysis using all achievement data for each student, even those with incomplete testing histories, to estimate the effects of teachers, schools and school systems. The development of this software has allowed the inherent advantages of longitudinal analyses to be extended to a statewide application, previously unavailable from commercial software. Compared to simpler approaches to educational value-added assessment, the SAS EVAAS system offers a number of advantages.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 201-271). San Diego, CA: Academic Press.

Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology*, *83*(4), 947-961.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and

physical attractiveness. *Journal of Personality and Social Psychology*, *64*, 431-441.

Benjamin, D. J., & Shapiro, J. M. (2009). Thin-slice forecasts of gubernatorial elections. *Review of Economics and Statistics*, *91*(3), 523-536.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328-375). New York, NY: Macmillan.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York, NY: Guilford Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: ASCD.

Downey, C. J., Steffy, B. E., English, F. W., Frase, L. E., & Poston, W. K., Jr. (2004). *The three-minute classroom walkthrough: Changing school supervisory practice one teacher at a time*. Thousand Oaks, CA: Corwin Press.

Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist*, *49*(8), 709-724.

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, *107*(1), 186-213.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25-42.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, *79*(4), 79-107.

Goldhaber, D. (2002). The mystery of good teaching: Surveying the evidence on student achievement and teachers' characteristics. *Education Next*, *2*(1), 50-55.

Goldhaber, D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129-145.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*(1), 84-117.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (NBER Working Paper 11154). Cambridge, MA: National Bureau of Economic Research.

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101-136.

James, W. (1950). *The principles of psychology*. New York, NY: Dover. (Original work published 1890)

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Kahneman, D. (2002, December 8). *Maps of bounded rationality: A perspective on intuitive judgment and choice* [Nobel Prize lecture]. Retrieved from http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahnemann-lecture.pdf

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases* (pp. 49-81). New York, NY: Cambridge University Press.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267-293). Cambridge, England: Cambridge University Press.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data* (NBER Working Paper 15803). Cambridge, MA: National Bureau of Economic Research.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54-78.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480-498.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.

McCaffrey, D. F., Han, B., & Lockwood, J. R. (2009). Turning student test scores into teacher compensation systems. In M. G. Springer (Ed.), *Performance incentives: Their growing impact on American K-12 education* (pp. 113-147). Washington, DC: Brookings Institution.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research*, *80*(4), 242-247.

Milanowski, A. (2004). The relation between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53.

Naftulin, D., Ware, J., & Donnelly, F. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, *48*, 630-635.

Neisser, U. (1963). The multiplicity of thought. *British Journal of Psychology*, *54*(1), 1-14.

No Child Left Behind Act, 20 U.S.C. 70 § 6301 *et seq.* (2002).

Nuthall, G., & Alton-Lee, A. (1990). Research on teaching and learning: Thirty years of change. *Elementary School Journal*, *90*, 546-570.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-257.

Ornstein, A. C. (1995a). Beyond effective teaching. *Peabody Journal of Education*, *70*(2), 2-33.

Ornstein, A. C. (1995b). The new paradigm in research on teaching. *Educational Forum*, *59*, 124-129.

Peterson, K. D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal*, *24*(2), 311-317.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Piaget, J. (1926). *The language and thought of the child*. London, England: Routledge Kegan Paul.

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS)*. Baltimore, MD: Paul H. Brookes.

Radmacher, S. A., & Martin, D. J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *Journal of Psychology*, *135*, 259-268.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology*, *133*(1), 19-35.

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Rosenshine, B. V. (1987). Explicit teaching. In D. C. Berliner & B. V. Rosenshine (Eds.), *Talks to teachers* (pp. 75-92). New York, NY: Random House.

Sanders, W. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education*, *14*(4), 329-339.

Sanders, W. L., & Horn, S. P. (1995). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. In A. J. Shrinkfietd & D. Stufflebeam (Eds.), *Teacher evaluation: Guide to effective practice* (pp. 337-350). Boston, MA: Kluwer.

SAS Institute. (n.d.). *Dr. William L. Sanders*. Retrieved from http://www.sas.com/govedu/edu/bio_sanders.html

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 470-428.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059-1074.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3-22.

Stallings, J. A., & Mohlman, G. G. (1988). Classroom observation techniques. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 469-474). Oxford, England: Pergamon.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, *23*(5), 645-665.

Strong, M. A. (2009). *Effective teacher induction and mentoring: Assessing the evidence*. New York, NY: Teachers College Press.

Vygotsky, L. S. (1987). Thinking and speech. In R. W. Rieber & A. S. Carton (Eds.), *The collected works of L. S. Vygotsky: Vol. 1. Problems of general psychology* (pp. 37-285). New York: Plenum Press. (Original work published 1934).

Ware, J., & Williams, R. (1975). The Dr. Fox effect: A study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education*, *40*, 149-156.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129-140.

Waxman, H. C. (1995). Classroom observations of effective teaching. In A. C. Ornstein (Ed.), *Teaching: Theory into practice* (pp. 76-93). Needham Heights, MA: Allyn & Bacon.

Williams, W. M., & Ceci, S. J. (1997). "How'm I doing?" *Change*, *29*, 13-24.

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

About the Authors

Michael Strong is a senior researcher at the Center for Research on the Teaching Profession at the University of California, Santa Cruz.

John Gargani heads Gargani + Company, Inc., an organization that conducts evaluation research in education and social settings.

Özge Hacifazlioğlu is an assistant professor in the Faculty of Arts and Sciences at Bahçeşehir University, Istanbul, Turkey.