SPECIAL ISSUE
Historical and Contemporary Perspectives on Educational Evaluation

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 26 Number 54 April 16, 2018 ISSN 1068-2341

Between Scylla and Charybdis: Reflections On and Problems Associated with the Evaluation of Teachers in an Era of Metrification

David C. Berliner
Arizona State University
United States

Citation: Berliner, D. C. (2018). Between Scylla and Charybdis: Reflections on and problems associated with the evaluation of teachers in an era of metrification. *Education Policy Analysis Archives, 26*(54). http://dx.doi.org/10.14507/epaa.26.3820 This article is part of the Special Issue, *Historical and Contemporary Perspectives on Educational Evaluation: Dialogues with the International Academy of Education,* guest edited by Lorin W. Anderson, Maria de Ibarrola, and D. C. Phillips.

Abstract: The Scylla and Charybdis in this discussion of teacher evaluation are standardized achievement test data on the one hand, and classroom observational systems on the other. These are the two most common methods used to judge teachers' competency. Both have serious flaws: the former primarily with validity, the latter primarily with reliability. At most these evaluation strategies provide teachers' and their supervisors information about which to converse. But these two methods have such serious flaws that they should never be used as the primary grounds for rewarding, punishing, or firing teachers. When both methods of evaluation are used to judge teacher competency, the correlation between achievement tests and observational data is quite low. When two methods claiming to assess the same construct do not correlate well, either one or both

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 1/15/2018
Revisions received: 2/18/2018
Accepted: 2/18/2018

methods are failing to assess the intended construct. There are two alternatives for navigating between Scylla and Charybdis: "Duties Based Teacher Evaluation" and "Performance Measures." These methods have much to recommend them, though like all methods of personnel evaluation, reliability and validity issues remain problematic.
Keywords: Teacher evaluation; bad teachers; standardized achievement tests; observational instruments; classroom observations; construct validation; duties-based teacher evaluation

Reflexiones sobre problemas asociados con la evaluación de maestros en una época de mediciones

Resumen: En esta discusión de la evaluación de los maestros Escila y Caribdis representan, por un lado, las pruebas estandarizadas de desempeño y, por el otro, los sistemas de observación de clase. Estos son los dos métodos más comunes para juzgar la competencia de los maestros. Ambos tienen serias deficiencias; el primero con su validez, el segundo básicamente con su confiabilidad. En el mejor de los casos, estas estrategias de evaluación ofrecen a los maestros y sus supervisores temas de conversación; pero ambos tienen serias deficiencias y no deberían usarse en ningún caso como fundamentos para premiar, castigar o despedir a los maestros. Cuando ambos métodos se usan para juzgar la competencia de los maestros, la correlación entre pruebas de desempeño y datos derivados de la observación es muy baja. Cuando los dos métodos aseguran evaluar el mismo constructo, no se correlacionan bien; uno o ambos fracasan al evaluar el constructo pretendido. Hay dos alternativas para navegar entre Escila y Caribdis: la "Evaluación basada en las obligaciones de los maestros" y las "medidas de desempeño". Estos métodos son muy recomendables, aunque como todo método de evaluación personal, la confiabilidad y la validez siguen siendo asuntos de discusión muy problemáticos.
Palabras-clave: Evaluación de maestros; malos maestros; pruebas estandarizadas de logro; instrumentos de observación; observación en clase; validación de constructos; evaluación basada en las obligaciones de los maestros

Reflexões sobre problemas associados à avaliação de professores em um momento de medições

Resumo: Nesta discussão da avaliação dos professores Scylla e Charybdis, eles representam, por um lado, os testes de desempenho padronizados e, por outro, os sistemas de observação de classe. Estes são os dois métodos mais comuns para julgar a competência dos professores. Ambos têm sérias deficiências; o primeiro com sua validade, o segundo basicamente com sua confiabilidade. Na melhor das hipóteses, essas estratégias de avaliação oferecem aos professores e seus supervisores tópicos de conversação; mas ambos têm sérias deficiências e não devem ser usados em nenhum caso como base para recompensar, punir ou demitir professores. Quando ambos os métodos são usados para julgar a competência dos professores, a correlação entre os testes de desempenho e os dados derivados da observação é muito baixa. Quando os dois métodos afirmam avaliar o mesmo constructo, eles não se correlacionam bem; um ou ambos falham em avaliar a construção pretendida. Existem duas alternativas para navegar entre Scylla e Charybdis: a "Avaliação baseada nas obrigações dos professores" e as "medidas de desempenho". Esses métodos são altamente recomendados, embora, como qualquer método de avaliação pessoal, confiabilidade e validade, ainda sejam questões muito problemáticas de discussão.

Palavras-chave: Avaliação de professores; maus professores; testes de realização padronizados; instrumentos de observação; observação em aula; validação de construtos; avaliação baseada nas obrigações dos professores

Between Scylla and Charibdis: Reflections on and Problems Associated with the Evaluation of Teachers in an Era of Metrification

In this article, I provide my views on the evaluation of teachers after 50 years of thinking about this issue as a parent, and as a professor of educational research. In the end, I stand with those teachers who protest government supported teacher evaluation systems based in whole or in part on standardized achievement tests that are used for high-stakes, highly consequential decisions about teachers. Certainly, the desire to have reliable and valid metrics for teacher evaluation is something we all share. But I am not sure if that is achievable, and, in my opinion, clearly isn't possible now. There are no teacher assessment systems that make use of data from standardized student achievement tests that I believe to be fair.

Standardized achievement tests for evaluating teachers are not fair because it is usually not the teachers that are most responsible for the poor performance of children on standardized achievement tests. Poverty (or wealth), and its sequelae, more than teacher competency, affects performance on those standardized tests. There is also a second prevalent approach to the evaluation of teachers: the use of classroom observation systems. These too can be unfair because they often suffer from unreliability. These two ways of assessing teacher quality place evaluators between Scylla and Charybdis: Neither approach works well.

Scylla was a female with 12 feet and six heads on long, snaky necks. Each head had a triple row of shark-like teeth. The loins of this most alluring lady were girdled by the heads of baying dogs. She lived on one side of the narrow passage between Sicily and the Italian boot. She would leave her cave to devour whatever sailing ships came within reach.

Another grand lady of the times, Charybdis, lurked under a fig tree on the opposite shore from Scylla. She drank down and belched forth the waters in that region three times a day. Thus, as the creator of whirlpools, she too was dangerous to the shipping in the region.

To navigate "between Scylla and Charybdis" means to avoid being caught between two equally unpleasant alternatives. In its more modern form it is to be caught between a rock and a hard place. Whether sailing in dangerous waters, or choosing between methods to evaluate teachers, choice can be difficult, and lives or careers can be threatened.

After discussing the problems inherent in both of these methods as a means of evaluating teachers, I conclude with a brief mention of two other forms of teacher evaluation that skirt some problems associated with assessment tests and observations. These are "duty based evaluation" and the evaluation of teacher competency by means of performance tests.

Why Evaluate Teachers?

Before we think further about how Scylla and Charybdis are apt descriptions of methods for appraising teachers, I want to note some differences about why we evaluate personnel in commerce and industry, and education. For example, in business we usually evaluate employees to decide on remuneration for the work being done, particularly if there have been changes in job duties and responsibility. Evaluations of this kind also help to decide bonuses, if the organization provides those for exemplary work. But for teachers, pay is often determined by a bureaucratic schedule, often related to years worked, degrees earned, and courses taken. And cash bonuses for teachers'

work are rarely given. When cash bonuses have been tried, they have usually been tied to student test scores. These clearly have not worked well in education (Amrein-Beardsley & Collins, 2012; Madaus, Ryan, Kelleghan, & Airasian, 1987). For the most part, teaching is a "flat" profession, with few opportunities to do much else than teach. So, reasons to engage in employee evaluation related to getting the compensation "right" for the kinds and quality of the duties performed are much less relevant to the teaching profession than for commerce and industry.

We also evaluate employees to determine the professional development that is needed by the staff of our businesses, so they can perform better at their jobs. This is especially true if changes are coming, such as new technology. Unfortunately, educational systems rarely have the money to provide teachers the professional development opportunities that might make teachers better, or prepare them for changes in curriculum and instruction. A good example of this problem in the U.S. context is the (ongoing) set of problems that are associated with the implementation of the relatively new Common Core State Standards (CCSS). While industry sometimes is willing to invest in preparing their employees for change, education typically does not do so. Thus, a good deal of the hostility to the CCSS has been generated by requiring major changes in curriculum and instruction, with little or no additional allocation of funds to prepare teachers for those changes.

Evaluating teachers in the USA is fundamentally different than evaluating personnel in commerce and industry. It is done *primarily* to get rid of "bad" teachers. It is this issue that concerns the public and teachers around the world. There is, of course, widespread agreement that our children must be protected from bad teachers. So, in the USA, no one argues about the necessity for teacher evaluation, and the right of a school district to dismiss bad teachers.

But how many "bad" teachers are there in the USA? Is there a reliable estimate of the base rate of our "bad" teachers? About four years ago I testified in a highly-publicized lawsuit about tenure rights in California. The judge asked me to estimate the percentage of "bad" teachers in the state. I made up an answer: "1, 2 or 3%!" This was based on my own classroom observations over many years.

I have continued to work on this issue since then and still have no reliable data to share. Nevertheless, my belief is that the base rate of bad teachers in the USA is remarkably low, while the system to identify them is too often costly, insensitive, and insulting. The belief that large numbers of American teachers are "bad," or put differently, that the base rate of bad teachers in the K-12 public school system is high, may be like the welfare queens that Ronald Regan talked about, the disability cheats that insurance companies talk about, and the fraudulent voters that our Republican congresspersons talk about. They simply may not exist in large numbers.

Estimates of the Percentage of "Bad" Teachers

In the ensuing four years I have asked the judge's question to hundreds of school administrators, school board members, and teachers. I set the question up this way: By "bad" I do not mean a teacher that is too strict or too permissive for your taste; or one that is using phonics while you believe in whole language, or vice versa; and I also don't mean a teacher that is temporarily having a bad time because of a divorce or illness; and I don't mean a teacher that isn't as sure of themselves in mathematics or science as we might want them to be. By a bad teacher I mean one who will hurt the children they teach. They will do this either by significantly retarding their progress, because the teacher has inadequate knowledge of what they teach; or they use methods, or hold attitudes that are harmful to some, or all of the children; or they have another job or difficult home life and cannot allocate the time needed to plan their classes adequately, nor muster the energy required to put in a proper days' work in a job that requires energy, empathy and continuous

attention. I ask my audiences, given their experiences, to estimate what percent of the teachers they have encountered are "bad" teachers, given the kind of loose (but reasonable) definition of a bad teacher that I just supplied?

From the hundreds of people to whom I have asked my question, I get a mean estimate of about 3%, with only a rare estimate over 7%. Charlotte Danielson (2007), the developer of the most popular instrument for observing and evaluating teachers, guesses that 6% of the many thousands of teachers that have been evaluated with her instrument are in need of remediation. The need for remediation, for Danielson, is related to performance that is below her standards for certain behavior. This is not the same as "bad." It is not unreasonable to assume that truly *bad* teachers, who fall into the category "needs remediation," could be half that rate of those who do need some forms of remediation. Peter Greene (2016), writing as the blogger "*curmudgeon*," thinks that Danielson's estimate is too high. And so he might also think that about 3%, or less is a reasonable estimate of the base rate of "bad" teachers.

For the child in the class of a bad teacher, and for that child's parent, it is little solace to learn that most teachers are not "bad" at all. We do need to keep in mind that the numbers of bad teachers, welfare queens, disability cheats, and fraudulent voters may all be products of our fears. Their base rates have not been determined by sound research and may be quite low.

Why might such a low base rate of "bad" teachers be an accurate estimate? First, it is not a random cross-section of citizens who become teachers. Declaring an education major in a well-regarded university usually requires a "reasonable" grade point average. In such institutions, a "B" or better, after two years of study, is the common grade point average required for entry into programs of teacher education. Because of that, the chance of getting a "incompetent" teacher is markedly reduced. However, this may not be the case in very small, or commercial and alternative teacher education programs. In some states, many of these small colleges with lower standards for entrance provide a substantial number of teachers for their states' schools.

Second, since the year 2000 there has been a steady climb in the number of teachers with SAT and ACT scores in the top third of those distributions. Roughly 40% of teacher education majors now come from the top third of those distributions, while fewer than 20% come from the bottom third (Goldhaber & Walch, 2014; Lankford, Loeb, McEachin, Miller & Wycoff, 2014). For a profession that is often disrespected, and with relatively low pay for the credentials required, education actually draws a much larger pool of talent than might be expected.

Third, most contemporary university programs are strongly clinical, or field based (American Association of Colleges for Teacher Education, 2010; Hammerness et al., 2005; National Council for Accreditation of Teacher Education, 2010). So, the chance of getting a teacher who has little or no experience in classrooms is considerably reduced. However, this is probably not true of commercial and proprietary teacher education programs, whose numbers have swelled because of the current teacher shortage. And it is certainly not true of the most of the teachers who come from the Teach for America program (Veltri, 2010).

Fourth, in our program of teacher education at Arizona State University, when we had full enrollment, we counseled out (removed) about 10% of the teachers whom we had initially let into the program. What this is likely to do, of course, is to reduce the likelihood of getting a bad teacher. In the past, this rate of dropping students was not unusual for teacher education programs at good universities. [However, the recent decline in candidates for teacher education programs and the current concerns about a shortage of teachers makes it likely that there will be less stringent oversight of trainees and novice teachers.]

Fifth, in the first few years of a novice teachers' career, principals and other district and school personnel counsel out, or fail to rehire, a substantial number of what they perceive to be

"bad" teachers. They only do this, however, when labor is available to staff all their classrooms. Principals I have interviewed say they would rather keep a marginal teacher than have no teacher at all to staff a class at the start of the year. The current shortage of teachers in the USA suggests that more marginal teachers will be retained, perhaps even tenured, then would be the case were there a more adequate supply of teachers.

Other novice teachers who feel unsuccessful, and those who learn that they do not enjoy classroom life, also leave the profession in the first five years. This too reduces the numbers of those who might eventually be labeled a "bad teacher." The rates of leaving or being removed from the profession in the first year is about 10%, and cumulatively, by year four, it is 17% (Gray & Taie, 2015). But these data were obtained *during* the recent recession. Before the recession, when jobs were much more plentiful, the rate of teachers' leaving the profession in the first five years, for any reasons, was about 40-50% (Di Carlo, 2011; Ingersoll, 2003).

Whatever the rate, existing evidence indicates that a higher percentage of those who left teaching were less effective than those instructors that stayed (Boyd, Grossman, Lankford, Loeb, & Wykoff, 2009). This also reduces the number of bad teachers in America's classrooms.

Base Rates of "Bad Professionals" in Other Professions

Are the rates of bad professionals in other fields likely to be the same as in education? That is hard to tell. But in medicine it was recently found that 1% of physicians accounted for 32% of paid malpractice claims over the past 10 years (Studdert et al., 2016). This indicates a small number of "bad" physicians. In a different study, by Public Citizen, one MD, Physician No. 33041, had at least 31 malpractice payments made on his behalf between 1993 and 2005, totaling more than $10 million in damages. So the malpractice rate, indicating large numbers of "bad" physicians, is quite low, although the damage they can do is substantial, and literally, sometimes, deadly. But the key finding here, is that the "bad" physician rate seems low. Sadly, so are the numbers who lose their license because of incompetence. While the public worries about bad teachers who are allowed to continue in their jobs, we have evidence that physicians found to be incompetent multiple times, are frequently keeping their jobs. And they can do a lot more damage.

When it comes to the legal profession we see a similar phenomenon. California has about 190,000 practicing lawyers (State Bar of California, 2017). In 2016, their ethics board received about 15,000 complaints about attorneys. This is an annual rate of unhappy clients of about 8%. But about 13,000 of these complaints were judged to be complaints without enough merit to be concerned about "bad" or "unethical" attorney behavior. As in education, and in medicine, many complaints in law are proffered, but whether a client's unhappiness reaches a level to warrant a charge of incompetence is quite a separate matter. Thus, the California bar filed complaints against only 672 lawyers, resulting in 444 disbarments, suggesting the annual rate of finding genuinely incompetent lawyers is less than 1%.

In the USA, whether we talk of social workers, nurses, physicians, lawyers or teachers, we are identifying individuals who enter their fields not only to be successful, but make a positive difference in the lives of others! Thus, it might well be expected that the rates of incompetence and unethical behavior among such morally committed and dedicated professionals is actually remarkably low. We know such behavior occurs in education. We repeatedly learn about teachers who cheat in testing, or inappropriately have physical contact with a student, or display biased behavior toward some group of students. But if the base rates in education and these other fields these fields are actually low, we need to be sure that the system is able to identify the few incompetent educational, medical, and legal professionals without destroying the professional lives of others in that profession. There seems to be a "search and destroy" policy to find the

incompetents that is hurting the huge numbers of hard working dedicated and competent professionals in education, medicine, and in other fields.

Danielle Ofri M.D., Ph.D., writing in the *New England Journal of Medicine* (2010), remarks that

> "Quantitative analysts will see it as a sign of medical arrogance that physicians insist that everyone simply trust us to do the right thing because we are such smart and noble people. I've always wanted to ask these analysts how they choose a physician for their sick child or ailing parent. Do they go online and look up doctors' glycated hemoglobin stats? Do they consult a magazine's Best Doctor listing? Or do they ask friends and family to recommend a doctor they trust? That trust relies on a host of variables — experience, judgment, thoughtfulness, ethics, intelligence, diligence, compassion, perspective — that are entirely lost in current quality measures (of physicians and nurses). These difficult-to-measure traits generally turn out to be the critical components in patient care."

I think Dr. Ofri is right. Experience, judgment, thoughtfulness, ethics, intelligence, diligence, compassion, perspective, and many other attributes like these, are the hallmarks of good professional practice in medicine as well as in education. But neither in medicine nor education can these attributes be measured reliably.

So, we start this look at the evaluation of teachers with two cautions. First, the base rate of bad teachers in the USA may be very low, and the reasons for that are quite sensible. I should note however, that the judge in the trial I mentioned earlier, said that if 3% of California's (roughly) 250,000 teachers were, indeed, "bad," that would mean that 7,500 "bad" teachers exist, and so tenure laws should be done away with, because, said the judge, tenure can too easily protect bad teachers.

A different way to look at these same data, if one accepts my totally made up figure of 3% bad teachers, is that California can claim their system is so remarkably good that 97% of California's teachers are adequate, or excel at what they do! That may actually be the case! But that idea is hard to sell to an angry parent convinced that their child is with one of the other kind of teachers. It is worth noting, too, that the judge was overruled by a higher court, though legal disputes about this issue are ongoing.

The second caution is that the characteristics that make for the kind of professional behavior we admire in physicians, nurses, lawyers and teachers are often quite hard, perhaps impossible to measure reliably. When we turn to more reliable measures for assessing characteristics of their professional competence, we may find that those more reliable instruments are less valid for determining the competencies of the professionals we are trying to evaluate. As mentioned earlier, the two major quantitative approaches to assessing and evaluating teachers are by means of standardized achievement tests (Scylla) and with classroom observational instruments (Charybdis).

Scylla: What are the Problems with Using Standardized Achievement tests for Evaluating the Competency of Teachers?

I have argued elsewhere (Berliner, 2014, 2015) that standardized achievement tests have numerous problems, especially when used in Value Added Models of evaluation (VAMs). They simply should not be used to evaluate teacher competency. Let me share just a few of these problems.

First, and foremost, is that the American Statistical Association (2014) has found that only between 1% and 14 % of the variance in standardized achievement tests can be attributed to the

teacher. So, the most important reason not to use a standardized achievement test is that it barely measures the teachers' effects on students. One of our finest scholars of measurement, Ed Haertel (2013), posits that on VAMs – where two standardized achievement tests are given, say, a year apart, on average, you can expect teachers to account for only about 10% of the variance in these tests. He argues that, on average, outside-of-school, and school factors that are outside of the classroom, are likely to influence 70% of the variance of these tests! What might some of these influences be? Inadequate medical, dental and vision care in family and neighborhood; percent of low birth-weight children in the neighborhood; food insecurity in the family; environmental pollutants in home and neighborhood; family relations and family stress; percent of mothers at the school site that are single and/or teens and /or do not possess a high school degree; language spoken at home; family income; mobility rates of families in the neighborhood; unavailability of high quality early education, and on and on. Other factors affecting the standardized achievement test scores, but also not under the teachers' control, include factors such as class size, teacher turnover or school churn rates, quality and frequency of professional development opportunities, availability of counseling and special education services for students, availability of librarians and school nurses, level of parent involvement, and on and on.

If you think like a politician or parent, it seems difficult to accept the idea that teachers do not affect standardized achievement test scores much at all. But think about it this way: suppose we give a fourth-grade standardized achievement test and then, a year later, we give a fifth-grade standardized achievement test to the same elementary school children. We do this to measure the "value added" by the fifth-grade teacher to the students' already impressive set of achievements. The fourth-grade standardized achievement test scores will correlate with the fifth-grade standardized achievement test scores at about .7 or better. The square of that is about .5, indicating that 50% of the variance in the second test, the one we might want to use to judge the value added by a teacher, is already accounted for by the teachers this child has had in past years, along with family social class and the opportunities for learning and development that social class confers. So half the variance we might want to attribute to a teacher is already accounted for.

Additionally, it is likely that the second test has some error in it, as all social science measures do, and that will account for about 10% more of the variance in the fifth-grade tests. Now only 40% of the variance is left to be accounted for, and of course this year's family events, which might include such things as illness, deaths, births, divorce, or job loss, will influence the scores on this years' tests as well. Then there are, as noted above, the many community events that might influence standardized achievement test scores during the year the child goes from fourth to fifth grade, including such things as flu epidemics and shootings. On top of that there are school events that influence achievement in a particular year, like the churn or stability of teachers, the firing or addition of librarians and counselors to the school staff, class size reductions or additions, and even the number of girls in the class. (In fact, the latter is quite reliably found to be a predictor of test scores, with more girls equaling higher scores. Moreover, this source of variance seems difficult to remove statistically (cf. Newton, Darling-Hammond, Haertel, & Thomas, 2010).

What all this means for the fifth-grade teacher who is being assessed and evaluated, whose value added to their students' total knowledge and skill is what we want to estimate, is that the variance in standardized achievement tests that is left over to be attributable to that teacher, is minimal (cf. Fantuzzo, LeBoeuf & Rouse, 2014). Scylla is a force to be reckoned with; she destroys methods of evaluation as well as ships.

Additionally, making it hard to judge a teachers' competency with a standardized achievement test is the fact that not a single standardized achievement test has ever shown that its items are instructionally sensitive. Imagine that some of the items on a standardized achievement

test are appropriate for a particular unit of instruction. Imagine further that this unit of instruction is taught by the best teacher in the state. Would the passing rate of these items from the standardized achievement test increase over what it was determined to be in the tryouts of the assessment? The question is whether the test items are actually reactive to good instruction? If we want to judge teachers' competency, we must have a measure that is sensitive to instruction, or the inference about a teachers' instructional competency cannot be justified. Currently we have no way of knowing if we do, or do not, have items that are reacting to instruction. No test developer has ever checked. None.

Using standardized achievement tests to judge teacher competency also sets the conditions for Campbell's law to come into play (1975): "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." Thus, we can expect gaming of the evaluation system used, and even cheating by teachers and administrators to get the scores they need to be judged competent, especially if they could get fired or earn bonuses (Nichols & Berliner, 2007).

Moreover, in systems using standardized achievement test scores to judge competency, teachers may too easily confuse successful teaching with good teaching. "Successful" teaching is about obtaining high test scores, say through excessive test preparation. On the other hand, "good" teaching—use of debates, small group work, project based learning, and so forth—may be sacrificed for the higher test scores that are required to keep one's job or receive a bonus.

There are many other reasons that standardized achievement tests cannot be used to evaluate teachers validly (Berliner, 2015). I think that standardized achievement tests have only two advantages. One is that they appear logically to be related to teacher effectiveness. So, the public, the media, and politicians like to use them, even if the vast majority of the research community tells them they cannot validly make the inferences they want, from the date they obtain.

The second major advantage of these tests is that they are remarkably cheap to use. The data are already collected as part of the accountability systems used in states and districts to judge student competency. So it seems sensible to just pay a little more for further analysis of the existing data, and turn the scores into VAMs of one kind or another to judge teachers, as well as students. What most who support this apparently sensible idea do not know is that a test designed to be valid for one purpose (assessing students) may not be valid for any other purposes (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

It should be recognized, however, that there is strong support from parents and policy makers for the regular assessment of achievement with standardized tests. Against the standards that have been created to guide learning at particular certain grade levels well-designed achievement tests do give insight into the performance of students and the schools they attend. Such tests are a direct measure of what the public expects the schools to accomplish. My concerns are about the sources of influence on those test scores, particularly about the amount of influence that teachers have on the scores obtained.

In the end, it appears to me that the most important factor preventing us from using better methods to assess America's teachers is the cost of evaluation. America's citizens want teachers to be evaluated, just as they want the potholes in the roads they travel to be fixed. But they don't want to pay very much for either.

Charybdis: What are the Problems with Using Observational Systems for Evaluating the Competency of Teachers?

Standardized achievement tests are indirect and distal measures of teacher competency. Observational systems are direct and proximal measures of teacher competency. Thus, observational measures have the potential of being more valid measures for the evaluation of teachers. There are many observational instruments in the USA, but two are particularly admired. One is the CLASS (Classroom Assessment Scoring System; see Table 1).

Table 1
Classroom Assessment Scoring System

Domain	Dimension	Dimension Description
Emotional Support	Positive Climate	reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions
	Teacher Sensitivity	reflects the teacher's responsiveness to the academic and social/emotional needs and developmental levels of individual students and the entire class, and the way these factors impact students' classroom experiences
	Regard for Adolescent Perspectives	focuses on the extent to which the teacher is able to meet and capitalize on the social and developmental needs and goals of adolescents by providing opportunities for student autonomy and leadership; also considered are the extent to which student ideas and opinions are valued and content is made useful and relevant to adolescents
Classroom Organization	Negative Climate	reflects the overall level of negativity among teachers and students in the class; frequency, quality, and intensity of teacher and student negativity are important to observe
	Behavior Management	encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior
	Productivity	considers how well the teacher manages time and routines so that instructional time is maximized; captures the degree to which instructional time is effectively managed and down time for students is minimized; it is not a code about student engagement or about the quality of instruction or activities
Instructional Support	Instructional Learning Formats	focuses on the ways in which the teacher maximizes student engagement in learning through clear presentation of material, active facilitation, and the provision of interesting and engaging lessons and materials
	Content Understanding	refers to both the depth of lesson content and the approaches used to help students comprehend the framework, key ideas, and procedures in an academic discipline; at a high level, refers to interactions among the teacher and students that lead to an integrated understanding of facts, skills, concepts, and principles
	Analysis and Problem Solving	assesses the degree to which the teacher facilitates students' use of higher level thinking skills, such as analysis, problem solving, reasoning, and creation through the application of knowledge and skills; opportunities for demonstrating metacognition (i.e., thinking about thinking), also included
	Quality of Feedback	assesses the degree to which feedback expands and extends learning and understanding and encourages student participation; in secondary classrooms, significant feedback may also be provided by peers; regardless of the source, focus here should be on the nature of the feedback provided and the extent to which it "pushes" learning

Table 1 cont.
Classroom Assessment Scoring System

Domain	Dimension	Dimension Description
Emotional Support	Positive Climate	reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions
	Teacher Sensitivity	reflects the teacher's responsiveness to the academic and social/emotional needs and developmental levels of individual students and the entire class, and the way these factors impact students' classroom experiences
	Regard for Adolescent Perspectives	focuses on the extent to which the teacher is able to meet and capitalize on the social and developmental needs and goals of adolescents by providing opportunities for student autonomy and leadership; also considered are the extent to which student ideas and opinions are valued and content is made useful and relevant to adolescents
Classroom Organization	Negative Climate	reflects the overall level of negativity among teachers and students in the class; frequency, quality, and intensity of teacher and student negativity are important to observe
	Behavior Management	encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior
	Productivity	considers how well the teacher manages time and routines so that instructional time is maximized; captures the degree to which instructional time is effectively managed and down time for students is minimized; it is not a code about student engagement or about the quality of instruction or activities
Instructional Support	Instructional Learning Formats	focuses on the ways in which the teacher maximizes student engagement in learning through clear presentation of material, active facilitation, and the provision of interesting and engaging lessons and materials
	Content Understanding	refers to both the depth of lesson content and the approaches used to help students comprehend the framework, key ideas, and procedures in an academic discipline; at a high level, refers to interactions among the teacher and students that lead to an integrated understanding of facts, skills, concepts, and principles
	Analysis and Problem Solving	assesses the degree to which the teacher facilitates students' use of higher level thinking skills, such as analysis, problem solving, reasoning, and creation through the application of knowledge and skills; opportunities for demonstrating metacognition (i.e., thinking about thinking), also included
	Quality of Feedback	assesses the degree to which feedback expands and extends learning and understanding and encourages student participation; in secondary classrooms, significant feedback may also be provided by peers; regardless of the source, focus here should be on the nature of the feedback provided and the extent to which it "pushes" learning

CLASS is a multidimensional framework that codes for quality indicators along 10 dimensions of effective classroom interaction, and then aggregates those into three quite reasonable domains—emotional support, classroom organization, and instructional support (Pianta, La Paro, & Hamre, 2008). It was developed for pre-K and other classrooms serving young students, but the instrumentation has been expanded and now covers pre-K through high school. It has a lengthy history of use, many admirers, and some solid research support (Gitomer et al., 2014).

A second instrument, used even more frequently in staff development and research, was developed by Charlotte Danielson (2007) and is called the Framework for Teaching (FFT; see Figure 1) The FFT is based on a constructivist model of teaching and requires observations in four domains of a teachers' professional life. These two instruments, and others, do get sufficient levels of inter-rater reliability from their raters, after a good deal of training in coding the behaviors of

interest. And the data obtained does show low correlations with student achievement, the most valued of the outcomes in education. This is all good. But Charybdis is out there waiting to sink ships and observation instruments alike.

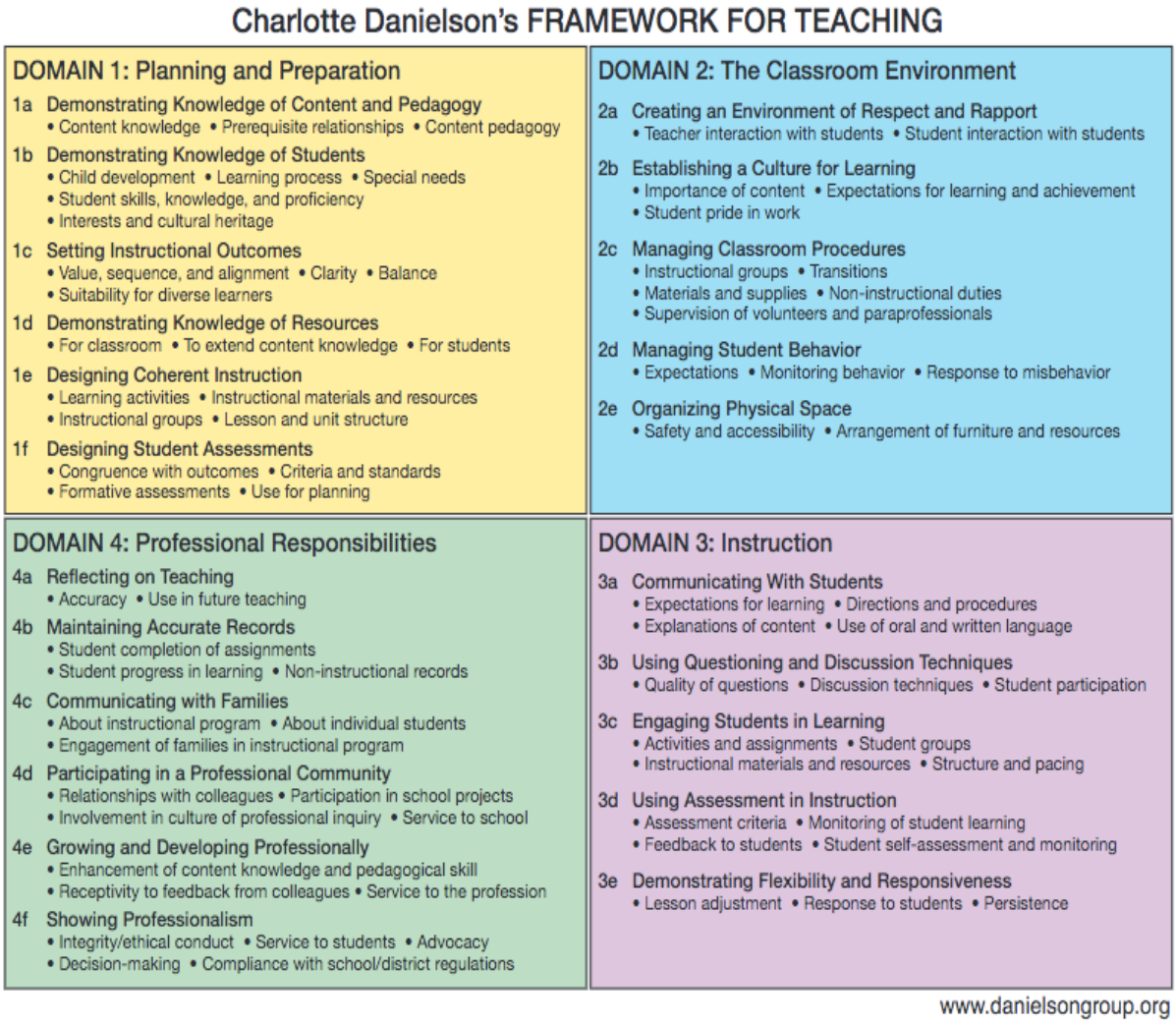

Figure 1. Charlotte Danielson's Framework for Teaching

First, and simply put, if the construct we are interested in is the effectiveness of the teacher in having students learn a designated curriculum, then each of these measures and the test scores we obtain from students ought to be moderately correlated. Both the achievement tests and the observation instruments claim to be measuring some aspect of that construct we call adequate or effective or good or excellent teaching. But, in fact, these observational instruments and tests of achievement are not correlated with each other very highly at all.

In the multi-million dollar MET study, funded by the Bill and Melinda Gates Foundation (Kane, McCaffrey, Miller, & Staiger, 2013) four of the observation instruments were correlated with the VAMs derived from math achievement test scores. Those correlations were .12, .18, .25, and .34. With the reading/language arts VAMs, three of the observation instruments correlated .12, .11, and .09. The variance in common is the square of those coefficients, indicating that they may not be measuring similar constructs at all, despite their claims. One or both of the constructs having to do

with effectiveness, as measured by means of the observations or the assessments, is not well represented.

In another sub-study of observation instruments and standardized tests, also done under the auspices of the Gates foundation, a special language arts observational instrument was correlated with the nationally standardized achievement test called the SAT 9, as well as the achievement test appropriate for the state in which that study was done. The two correlations between an observational measure of excellence in teaching, and both measures of excellence in teaching derived from VAMs, were .16 and .09. (Grossman, Cohen, Ronfeldt, & Brown, 2014). In a recent study by Strunk, Weinstein, & Makkonen (2014) the correlations between observational data and VAMs for reading and math, over one year, were .216 and .178. When the VAMs were accumulated over three years to have a more reliable indicator of teacher competency (because single year VAMs are not very reliable), the correlations turned out to be even lower (about .14 in both reading and in mathematics). The variance held in common in the measures of teacher competency via observational instruments and via standardized achievement tests was under 5% in all four analyses undertaken. Another recent study by Morgan, Hodge, Trepinski, and Anderson (2014), found correlations between observations and tests that were roughly between .20 and .40, indicating that these two different measures of determining exemplary teachers only have in common between 4% and 16% of the variance observed. These investigators noted that neither teacher performance in classrooms, nor teacher effectiveness as judged by test scores, were highly stable over multiple years of the study.

So, we have a conundrum: the criterion by which we judge the validity of our observations is often a standardized achievement test. And the criterion by which we can judge the validity of the standardized tests are often some kind of classroom observation instruments. But these two types of instruments share little variance in common. The observational scores and the standardized test scores almost always correlate under .30, and thus only about 10% of the variance is shared. This is not a reassuring state of affairs.

While I favor the evaluation of teachers through observation methods, rather than by means of any standardized achievement tests, most who engage in observational analyses of teaching forget that Charybdis is out there waiting to wreck such systems. The common problem with observation systems, and perhaps contributing to their low correlation with achievement tests, is not the unreliability of raters or coders. This can usually be managed though extensive training. It is, instead, the stability of teacher behavior in particular contexts that is being rated or coded. This is a hidden problem: Charybdis is sneaky as well as difficult to get around. Let me explain.

About 40 years ago Richard Shavelson was commissioned by me to run an observational study using generalizability (g) theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Erlich and Shavelson 1978). The project I was working on needed help in figuring out how many observers, and how many observations we would need to reliably code teacher behavior that we thought to be important, and for which we had designed an observational system. In what has to be one of the most ignored studies in the history of research on teaching, Shavelson found that only one observer visiting a classroom on one occasion can reliably code only a few behaviors. These easy to code behaviors are usually "high inference" variables, such as a rating of the teachers' enthusiasm, orderliness, preparedness, and other "trait-like" characteristics of teachers-- behaviors that are likely to persist throughout the day, and also from day to day. These are not unimportant teacher characteristics, and surely many of these traits of teachers are related to our notions of quality in classroom teaching. But even for these high inference variables there may be problems. Calkins Borich, Pascone, Kluge & Marston (1997) showed that for 12 teachers, measured three times by

different raters, almost 70% of the high inference variables they were interested in were unstable or unreliable.

Unreliability, however, is even more frequently found for less trait-like variables, those, for example, that are of interest when we try to code classroom interaction. For many of these more molecular behaviors, those that are more "state-like" variables than they are "trait-like" variables, the research suggests a rule of thumb. It may be that someone would need five or more observations, and multiple or extremely well trained observers, in order to reliably estimate the frequency and quality of many of the behaviors of the teacher, the student, or those emanating out of the teacher-student interaction.

These findings move observation instruments into the clutches of Charybdis. For example, the well-known and often used Teacher-Child Dyadic Interaction system of Brophy and Evertson (1976), the basis of dozens of research studies, allows for 167 variables to be coded. But generalizability (G) theory reveals that only 35 of these were found to have the necessary reliability from which to draw valid inferences (Erlich & Borich, 1979).

We can think about the problem this way. It might well be that "no response" to a teacher's question is an important behavior to code when observing classrooms. Theoretically, high rates of "no response" to questions when coding classroom interaction might indicate teachers who cannot ask "good," "well formed," "germane" questions of their students, or that student reticence to answer is because wrong answers to a teacher questions is often met with ridicule. That is worth knowing. It could also mean that that students were not prepared to answer the questions that were asked by the teacher. That too is worth knowing. Either way, "no response" to teachers' questions has implications for evaluating a teacher, and perhaps that information can be used for the design of staff development, as well. But it appears that nine occasions of coding, lasting at least three hours each, are needed to get the reliability of the measure of "no response to teachers' question" up to .70, a level of reliability sufficient for making inferences about a teachers' behavior (Erlich & Borich, 1979).

Let us examine two other coding categories. Suppose we posit that a teacher's response to a student's wrong answer to a question is a teaching skill we believe to be important? If so, it might take between five and eight occasions to reliably code the various responses the teacher might make to students' wrong answers. Do teachers' overreactions to student misbehavior indicate a problem to be corrected? If so, it is likely to take seven occasions to record this behavior reliably.

Pretorius et al. (2014) studied five lessons of 38 teachers and examined what they called "cognitive activation"—the evoking of students' thinking skills by teachers. This is, of course, the skill considered as the most important 21st century skill for the work force of the future to possess. And it is the skill that the PISA tests set out to measure every three years. But to get a reliable handle on this skill of teachers would likely require nine occasions before we can reliably differentiate between teachers who are good and those who are not good at cognitive activation.

In an influential paper in this area, Shavelson and Dempsey-Atwood (1976) looked at dozens of studies to determine the generalizability, that is, the stability of the observations made in those studies. They concluded that most studies using observations of classrooms are methodologically inadequate, that stability of teacher behavior is not found frequently enough, and that our measures are, therefore, too often unreliable. And here is their most important conclusion of all: They say that neither improved measurement nor new conceptualizations will fix the problem.

The reason for their negativism and mine is simple, but hard to accept by those who want a more stable and predictable world. Systematic variation is lacking for most of the teaching behaviors that we want to observe. Teachers, to be effective, must constantly monitor and change their behavior: they must adapt to subtle clues about changes in the instructional milieu. On the other

hand, unsystematic variation, giving us wobbly and unstable variables to examine, commonly occurs. This is because of the myriad subtle but powerful factors that make teaching so complex. Observations of life in classrooms are affected by the place a class is in during a particular unit of instruction (beginning, middle, or end of the unit); the observations are affected by the mood of the classroom on a particular day; observations are affected by events in the personal life of the teacher; they are affected by the time of day and the time of year; they are affected by who is absent and who is present on the day of observation; they are affected by whether the teacher has a sick baby at home, or a spouse who is drinking, and so forth. Even the weather affects what is observed!

So, the bottom line is this. If trained raters who are also practicing teachers would observe two teachers a day, the costs might be around $500 per observation, about $1,000 per day.[1] Thus, to find out if one particular teacher can or cannot ask "answerable" questions, we would likely require about nine observations, at about $500 per day, or $4,500, to get that piece of information reliably. On the other hand, if we intend to use one observer on one occasion to gather information, as we so often do, we can do so at much less cost and meet our obligation to evaluate teachers, *even if a good deal of what was coded and rated and judged is unreliable*.

To get reliable information about some apparently important teacher behaviors using observational techniques clearly costs too much money. Yet the assessment of teaching using standardized achievement tests raises validity problems. So, those who use the two most common methods for the evaluation of teachers—student achievement tests and observation instruments--are located between Scylla and Charybdis. They are caught between a rock and a hard place if they want to make consequential decisions about teachers. Neither form of evaluation is inappropriate to hold conversations with teachers about their students' performance, or their own. It is when consequential decisions are made from data derived from either source that serious ethical problems arise.

A personal note: When in doubt about which of these measures to trust more, I personally would always choose a direct and proximal measure of teacher competence, instead of an indirect and distal measure of competence. I trust the observations and evaluations of classroom artifacts by trained board-certified teachers and principals, with or without formal observation instruments, warts and all. Mere classroom visits, for short periods of time, by untrained observers are not what I have in mind.

A complex act like teaching, performed for six hours a day over 180 days, simply may not yield easily to quantification and metrification, despite our fondest hopes. In this age of metrification we need to be aware that not everything that can be counted counts, and not everything that counts can be counted (Cameron, 1963).

Many years ago I rejected Elliott Eisner's (1976) ideas about educational evaluation being akin to connoisseurship. I was sure that achievement tests and observational methods could be found that worked in the ways we needed them to work for reliable and valid teacher evaluation to take place. But now, older and less sure about my youthful dreams of technocratic solutions to the problem of evaluating teachers fairly, Eisner's ideas have much more currency for me.

The essence of the construct we are trying to get a handle on, teacher competency, is elusive: it is a chimera, it is a will-o-the-wisp, and closer to the arguments about what is, or is not, bad, good, or great art. High quality teaching may not be anywhere as easily judged as ice skating, gymnastics, and high diving in the Olympics. And we should note that even there, they always use highly trained judges, they also use multiple judges from multiple countries, they often throw out the top and

[1] This is my estimate of the average teachers' salary and benefits per day, plus training costs, transportation to site, and the salary and benefits of a substitute teacher to cover the observing teachers' classes.

bottom scores obtained, and in many of the sports judged they have many heats and semi-finals to winnow down the field for the finals. These heats, or semi-finals, are another way of saying that they have multiple occasions to judge who are the best athletes and teams in a sport, and recent objective records of their competence as an athlete. Olympic judging of sport appears to be a more costly and a better model for judging athletic competency than we have for judging the competency of our teachers. It seems as if many nations have decided that it is more important to identify, train, and pay for developing a competitive high diver, for the honor of their country, than it is to pick, train and pay for the development of a good teacher, for the future of their country.

Skirting Scylla and Charybdis: Duty-based Teacher Evaluation and Performance Tests of Teachers

The Duties-Based Approach

I will mention only two ways to escape the monsters Scylla and Charybdis. The first of these was offered by Michael Scriven (1994), who noted that teachers have certain duties to perform, just as do physicians and nurses. He has provided an extensive list of these. And then he asks why we do not simply judge teachers on whether they fulfill the essential duties of their profession, much like the assessment practices in some other professional fields. An overview of the major categories of his much larger list of teacher duties is given as Figure 2.

The fulfilling of duties, say, grading papers and tests in a reasonable time, preparing visuals to accompany the teaching of hard topics, or helping younger teachers learn their skills, are necessary though not sufficient conditions for being an excellent teacher. They do however, present less reliability problems than more nuanced judgements, say, whether the feedback accompanying a returned test was appropriate. Or whether the visuals used to explain difficult concepts were any good. Assessing the fulfillment of the duties of teaching provides reason to believe we have identified an adequate teacher. On the other hand, not fulfilling the requisite duties of teaching puts the spotlight on teachers in need of remediation, or perhaps, even dismissal.

Scriven notes how trained, experienced evaluators, using a duties-based evaluation system for describing teachers' behavior, have many goals. These evaluations can help in the design of staff development, can inform teacher training institutions about some deficits they have, and perhaps most important of all in modern times, these evaluations can be used for summative purposes. Duties based teacher evaluations can assist personnel decision by principals, personnel officers, superintendents, or school boards, and they will stand up in a court of law, or at an arbitration hearing, when a personnel decision is appealed. Evaluating teachers in this way is much closer to the way some other professionals are evaluated. This system avoids the dangers posed by Scylla and Charybdis because it does not go along with the pretense of having "objective" quantitative evaluations of teachers. Duties-based assessments examine the presence or absence of those things required to do one's job. I find this approach to be of great interest.

DUTIES OF THE TEACHER

Michael Scriven

This checklist consists of the headings from a long analysis of this topic, consisting of the checklist plus explanatory text for each of the checkpoints (Journal of Personnel Evaluation in Education, 1994, vol. 8, no.2, pp.151-184). The checklist provides a good overview of the whole approach, however, and is based on a complex evaluative theory, which includes, for example, the ethical principle that one cannot evaluate teachers by looking at the teaching style they employ, except insofar as this is prescribed by the accepted duties of a teacher. They can use much or little lecturing, question asking, etc., no matter what the research shows, just so long as they successfully cause the acquisition of valuable knowledge, skills, and attitudes in the areas for which they are responsible, at a rate that is appropriate or better for comparable students, within current ethical, resource, and legal parameters. Teachers have no duty to teach using a particular style, only to teach successfully. It is weakly sequential because there are socio-political reasons for each item's placement; e.g., the main reason for placing item 1 first was the perceived climate for acceptance by school boards, state and federal agencies, and parents.

1. **KNOWLEDGE OF SUBJECT MATTER**
   - A. In the field(s) of appointment, e.g., middle school mathematics
   - B. In across-the-curriculum subjects, e.g., composition, spelling
2. **INSTRUCTIONAL COMPETENCE**
   - A. Communication skills (use of age-appropriate vocabulary, examples, inflection, body language)
   - B. Management skills
     - a. Management of (classroom) process, including discipline
     - b. Management of (individual student's educational) progress
     - c. Management of emergencies (fire, tornado, earthquake, flood, stroke, violent attack)
   - C. Course construction and improvement skills
     - a. Course planning
     - b. Selection and creation of materials
     - c. Use of special resources
       - i. Local sites
       - ii. Media
       - iii. Specialists
     - d. Evaluation of the course, teaching, materials, and curriculum
3. **ASSESSMENT COMPETENCE**
   - A. Knowledge about student assessment options
   - B. Test construction and administration skills
   - C. Grading, ranking, scoring practices
     - a. Process (doing it correctly, i.e., using scoring keys, blind scoring)
     - b. Output (the results meet appropriate standards, e.g., (usually) not all As or all Fs)
   - D. Recording and reporting student achievement
     - a. Knowledge about options and obligations in reporting achievement
     - b. Good reporting process (to students, administrators, parents, authorized others)
4. **PROFESSIONALISM**
   - A. Professional ethics
   - B. Professional attitude
   - C. Professional development
   - D. Service to the profession (some but not each of the following)
     - a. Knowledge about the profession
     - b. Helping beginners and peers
     - c. Working for professional organizations
     - d. Research on teaching
   - E. Knowledge of duties
   - F. Knowledge about the school and its community
5. **NONSTANDARD BUT CONTRACTUAL DUTIES**
   e.g., supervision of chapel services in a religious school

Figure 2. Duties of the Teacher

Performance Tests of Teaching

Performance test of teaching are the last of the major forms of teacher evaluation to be discussed. Fifty years ago Popham (1971), was designing performance tests of teaching and I was impressed with them then, as I am now. They too have some reliability and validity problems as all assessments do. And perhaps their greatest problem is that they are not actual measures of teaching competence. Instead, they are a proxy for the skills that are thought to be related to competence. In the 1980s Shulman and his students and colleagues (1987, 1988) worked on performance tests too. They were designing prototypes for the National Board for Professional Teaching Standards, about which I'll say more in a moment. Darling Hammond and her colleagues (Darling-Hammond, 2010; Pecheone & Chung, 2006) developed a performance assessment called PACT—Performance Assessment for California Teachers. PACT is a pre-service performance assessment that asks for a demonstration of a wide range of teaching skills. The test is taken at the end of fieldwork associated with teacher education coursework. Of special note is that scores on the PACT correlated quite a bit higher with student assessment data than have the observational measures I mentioned above. Thus, we may conclude that the constructs that are measured by this pre-service performance test, and the constructs measured by a test of student achievement given after teachers have been doing actual teaching, show modest overlap. The PACT has been turned into a national test called the edTPA, administered by a private corporation. It costs a candidate for a teaching position $300 to take. But since the test has some modest predictive validity, it is a way of hiring teachers more likely to succeed, and thus is a mechanism for keeping that base rate of bad teachers to 3% or less. A performance test like the edTPA serves the same purposes as the medical boards and the bar exam—it can signal what is important to know, and it can keep out of the profession those whose performance on the test is judged to be insufficient enough to join the profession.

Over the last 30 years or so in the USA we have developed the National Board for Professional Teaching Standards. That Board administers performance tests of teaching for a wide variety of subject areas in different grade levels. My own work on teacher expertise informed the design of these tests, as did Shulman's prototypes and Darling-Hammonds' work. I bring this system to your attention because one study of these performance tests makes the case for further design and use of this form of assessment for practicing teachers.

In brief, here is the study (Bond, Smith, Baker, & Hattie, 2000). Two samples of teachers were recruited from among those who had attempted to obtain National Board Certification in the areas of Middle Grade Level/Generalist, or Early Adolescent Level/English Language Arts. One of the comparison groups (N= 31) consisted of those who passed the National Board examinations, the other comparison group (N=34) consisted of those who did not achieve Board certification through the assessment test. All the teachers were well experienced, had prepared diligently for the examinations, and spent considerable amounts of money to demonstrate they were highly accomplished teachers. In advance of visiting the classrooms of these 65 teachers, 13 features of expert teachers were hypothesized and observation instruments were developed to look at each of these. Classroom observers were trained and were blind as to which class they were observing—a teacher who had, or a teacher who had not passed the performance test.

This was a little study run by advocates of the Boards' approach to testing, but the results are quite remarkable. The Board-certified teachers, in comparison to those who failed to meet the Board standards on the assessments, excelled on every prototypical feature of expertise in classroom instruction. When looked at as effect sizes, the differences between these two highly experienced and confident teacher groups, on the 13 behaviors being assessed, ranged from just over one-quarter of a standard deviation to 1.13 standard deviations in favor of the Board-certified teachers. Thus, teachers found to be expert on the basis of the assessments of the performance test were anywhere

from 8 to 37 percentile ranks higher on measures that rated their use of knowledge, the depth of their representations of knowledge, their expressed passion, their problem-solving skills, and so forth.

This study provides predictive validity for the performance assessment program designed to identify highly effective teachers. The authors claim they can "Identify… and certify… teachers that are producing students who differ in profound and important ways from those taught by less proficient teachers. These students appear to exhibit an understanding of concepts targeted in instruction that is more integrated, more coherent, and at a higher level of abstraction than understanding achieved by other students" (Bond, Smith, Baker & Hattie, 2000, p. 113).

In another study the test scores of 600,000 elementary students from North Carolina were examined over a three-year period by a research team unconnected with the National Board (Goldhaber & Anthony, 2007). They found that Board-certified Teachers were far more likely to improve student achievement on the state's standardized tests than non-Board-certified Teachers. Board-certified Teachers raised student achievement about 7% more on math and reading tests than did teachers who took the tests but failed to get certified. The Board-certified teachers had their greatest impact with younger and with low-income students, with the scores of these students up to 15% higher than the scores of students who did not have Board-certified Teachers.

One of my students (Vandevroot, 2004), also found effects for Board-certified teachers. The bottom line is that valid performance tests of teaching can be designed, *if money is spent to do so*. It costs a lot of money to design the edTPA and at least a few hundred thousand dollars to develop a valid performance test of teaching in *each* of the 30 or so different areas of teaching in which the National Board has invested. To sit for these tests, and have them reliably scored is also expensive—around $2,500 for each candidate that wants to sit for the National Board test. These fees are rarely covered by employers of the teachers who take the tests. The point, however, is that both the edTPA and the National Board performance tests are able to identify more and less effective teachers. If the costs were to ever to be acceptable, we would not have trouble identifying more and less competent teachers.

Conclusion

What do we know about the various forms of assessment for evaluating teachers? We know that Standardized Achievement Tests, especially as VAMs, are unreliable and invalid, but relatively cheap to use. Observational methods are rarely even moderately correlated with achievement test scores, and often only provide reliable information about important aspects of teaching if many more than one occasion is used to judge teachers' competency. This becomes very expensive v ery quickly. Currently observational instruments do exist for which observers can be trained to agree on what they code, but the question of the stability of the behavior that was coded over time and occasions is not adequately addressed.

In the observational category we can also place the classroom visits of highly trained connoisseurs. These aestheticians of educational processes, observers that themselves may have been regarded as master teachers, is not usually accepted as reliable and valid for making consequential decisions about the quality of practicing teachers. But teaching, like performance on a balance beam, has both technical and aesthetic elements. Who, then, is better to judge a performance in either domain than someone who themselves was a highly successful practitioner in that domain, a successful teacher or gymnast. But it is also true that this and *all other observation methods* are costly.

Duties-based teacher evaluation never seems to catch on, but it has much to recommend it. It is comparatively cheap, in part because single raters can be trained to use this technique. Further, complex aesthetic judgements are not required, and thus fewer visits to classrooms or schools may be required.

Finally, performance tests of teaching have much to recommend them when trying to identify exemplary and poor teachers. But if such tests are to be used for decision-making, their validity must be substantial. Valid performance tests cost a lot to develop, and therefore a lot to take,

In a more ideal world, for deficiencies in performance that might be found, by whatever means of teacher evaluation that is used, there would be available a pool of funds for professional development (though those that provide such opportunities also have problems with demonstrating effectiveness). Evaluations of any kind seem more likely to find cause for remediation than to uncover incompetence serious enough to justify the dismissal of a teacher. As noted above, the base rate of "bad" teachers is likely to be low. But because funds for teacher development are not often available to accompany teacher evaluations, the evaluations often lead to teacher cynicism about any of the evaluation systems that are used. This is because too many teachers found to be poorly performing are not given remediation, and as a consequence, the more competent teachers and the schools in which these teachers work have their reputations damaged.

In summary, choosing to evaluate teachers via achievement tests or with observational methods places evaluators between Scylla and Charybdis. The form these monsters take is by creating problems with unreliability, and with construct, predictive, and consequential validity. But both methods yield metrics, and in contemporary times such metrics are desired, even if they are often uninterpretable. Performance tests of teachers can be designed to avoid many of these problems, but if they are to be used for any consequential decisions, they are very expensive to develop. Thus, it may be that connoisseurship and duties-based evaluations of teachers might provide the only cost effective approaches to teacher evaluation that can avoid the monsters. But these are not forms of teacher evaluation accepted as appropriate by either our teachers or our political leaders. Thus, the evaluation of teachers is likely to remain a mess.

Acknowledgements

The author thanks the careful reading and feedback given by Professors Lorin Anderson and Richard Shavelson. They improved what was originally a speech into a much-improved paper. All faults remaining, however, are mine.

References

American Association of Colleges for Teacher Education. (2010). *The Clinical Preparation of Teachers: A Policy Brief*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Statistical Association. (2014, April 8). *ASA Statement on* Using Value-Added Models for Educational Assessment. Washington, DC: Author.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and Unintended Consequences. *Education Policy Analysis Archives, 20*(12). Retrieved December 23, 2012, from http://epaa.asu.edu/ojs/article/view/1096

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record, 116*(1). Retrieved July 21, 2014, from http://www.tcrecord.org/content.asp?contentid=17293

Berliner, D. C. (2015, August 11). Teacher evaluation and standardized tests: A Policy Fiasco. Paper presented at the meetings of the international Academy of Education meetings and Melbourne University, Melbourne Australia. Retrieved May 22, 2017, from http://education.unimelb.edu.au/news_and_activities/events/upcoming_events/dean_lecture_series/dls-past-2015/teacher-evaluation-and-standardised-tests-a-policy-fiasco

Bond, L., Smith, T., Baker, W., & Hattie, J. (2000). The Certification System of the National Board for Professional Teaching Standards: A Construct and Consequential Validity Study. University of North Carolina at Greensboro, Center for Educational Research and Evaluation.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2009). *"Who leaves?" Teacher attrition and student achievement*. Working Paper 23. National Center for Analysis of Longitudinal Data in Education Research. Washington DC: Urban Institute.

Brophy, J. E., & Evertson, C. M. (1976). *Learning from teaching:* A developmental perspective. Allyn and Bacon.

Calkins, D., Borich, G. D., Pascone, M., Kluge, S., & Marston, P. T. (1997). Generalizability of teacher behaviors across classroom observation systems. *Journal of Classroom Interaction, 13*, 9–22.

Cameron, W. B. (1963). *Informal sociology: A casual introduction to sociological thinking*. Volume 21, Studies in Sociology. New York: Random House.

Campbell, D. (1975). Assessing the impact of planned social change. In G. Lyons (Ed.), *Social Research and Public Policies: The Dartmouth/OECD Conference.* Hanover, NH: Public Affairs Center, Dartmouth College.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley.

Danielson, C. (2007). Enhancing Professional Practice: A Framework for *Teaching* (2nd ed.). Washington, DC: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2010). Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching. Washington, DC: Center for American Progress.

Di Carlo, M. (2011). *Do half of new teachers leave the profession within five years?* New York: Albert Shanker Institute. Retrieved May 4, 2017, from http://www.shankerinstitute.org/blog/do-half-new-teachers-leave-profession-within-five-years

Eisner, E. W. (1976). Educational connoisseurship and criticism: Their form and functions in educational evaluation. *Journal of Aesthetic Education, 10*(3/4), 135-150. Retrieved May 5, 2017, at http://www.jstor.org/stable/3332067

Erlich, O., & Borich, G. (1979). Occurrence and generalizability of scores on a classroom interaction instrument. *Journal of Educational Measurement, 16*(1), 11–18.

Erlich, O., & Shavelson, R. J. (1978). The search for correlations between measures of teacher behavior and student achievement: Measurement problem, conceptualization problem, or both? *Journal of Educational Measurement, 15*(2), 77-89. DOI: 10.1111/j.1745-3984.1978.tb00059.x

Fantuzzo, J. W., LeBoeuf, W. A., & Rouse, H. L. (2014). *An investigation of the*relations between school concentrations of student risk factors and student educational well- being. *Educational Researcher, 43*(1), 25–36.

Gitomer, D., Bell, C., Qi, Y. McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: lessons from a classroom observation protocol. *Teachers College Record, 116*(6), 873–881.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*(1), 134-150.

Goldhaber, D., & Walch, J. (2014). Academic capabilities of the U. S. teaching*force are on the rise. Education Next, 14*(1). Retrieved May 4, 2017, from http://educationnext.org/gains-in-teacher-quality/

Gray, L., & Taie, S. (2015). Public school teacher attrition and mobility in the first five years: Results from the first through fifth waves of the 2007–08 beginning teacher longitudinal study (NCES 2015-337). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved May 5, 2017, from http://nces.ed.gov/pubsearch.

Greene, P. (2016, April 19), *Curmuducation*. Retrieved May 8, 2017, from http://curmudgucation.blogspot.com/search?q=Danielson

Grossman, P., Cohen, J., Ronfeldt, M, & Brown, L (2014). The Test Matters: The Relationship between Classroom Observation Scores and Teacher Value-Added on Multiple Types of Assessment. *Educational Researcher, 43*(6), 293–303.

Haertel, E. H. (2013), Reliability and validity of inferences about teachers based *on student test scores.* The 14th William H. Angoff Memorial Lecture, Princeton, NJ: Educational Testing Service.

Hammerness, K., Darling-Hammond, L., Bransford, J., Berliner, D., Cochran-Smith, M., McDonald, M., & Zeichner, K. (2005). How teachers learn and develop. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 358-389). San Francisco: Jossey-Bass.

Ingersoll, R. (2003). Is there really a teacher shortage? Center for the Study of Teaching and Policy, University of Washington, Seattle. Retrieved May 4, 2017, from http://ctpweh.org/

Kane, T. J., McCaffrey, D. F., Miller, T. & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. MET Project Research Paper. Retrieved May 4, 2017, from http://k12education.gatesfoundation.org/resource/have-we-identified-effective-teachers-validating-measures-of-effective-teaching-using-random-assignment/Using

Lankford, H., Loeb, S., McEachin, A., Miller, L. C., & Wycoff, J. (2014).Who enters teaching? Encouraging evidence that the status of teaching is improving. *Educational Researcher 43*(9), 444-453.

Madaus, G. F., Ryan, J. P. Kelleghan, T., & Airasian, P. W. (1987). *The Irish Journal of Education / Iris Eireannach an Oideachais, 21*(2), 80-91.

Morgan, G. B., Hodge, K. J., Trepinksi, T. M., & Anderson, L. W. (2014). The stability of teacher performance and effectiveness: Implications for policies concerning teacher evaluation. *Education Policy Analysis Archives, 22*(95). http://dx.doi.org/10.14507/epaa.v22n95.2014

National Council for Accreditation of Teacher Education (NCATE). (2010). Transforming teacher education through clinical practice: A national strategy to prepare effective teachers. Washington, DC: Author.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). *Value* added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational*

Policy Analysis Archives, 18(23). Retrieved March 27, 2012, from: http://epaa.asu.edu/ojs/article/view/810

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Ofri, D. (2010, August 12). Quality Measures and the Individual Physician. *New England Journal of Medicine, 363*, 606-607. Retrieved May 3, 2017, from DOI: 10.1056/NEJMp1006298.

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

Popham, W. J. (1971). Performance Tests of Teaching Proficiency: Rationale, Development, and Validation. *American Educational Research Journal, 8*(1), 105-117. 10.3102/00028312008001105

Praetorius, A-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2-12. https://dx.doi.org/10.1016/j.learninstruc.2013.12.002

Scriven, M. (1994). Duties of the teacher. *Journal of Personnel Evaluation in Education, 8*(2), 151-184. http://dx.doi.org/10.1007/BF00972261

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46*(4), 553. http://dx.doi.org/10.3102/00346543046004553

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Shulman, L. S. (1987). Assessment for teaching: An initiative for the profession. *Phi Delta Kappan, 69*(1) 38-44.

Shulman, L. S., Haertel, E., & Bird, T. (1988). *Toward alternative assessments of teaching: A report of work in progress*. Stanford, CA: Stanford University, School of Education, Teacher Assessment Project.

State Bar of California. (2017). 2016 Annual Discipline Report. Retrieved May 5, 2017, from http://www.calbar.ca.gov/AboutUs/Reports.aspx

Strunk, K., Weinstein, T., & Makkonen, R. (2014). Sorting out the signal: Do multiple measures of teachers' effectiveness provide consistent information to teachers and principals? *Education Policy Analysis Archives, 22*. http://dx.doi.org/10.14507/epaa.v22.1590

Studdert, D. M., Bismark, M. B., Mello, M. M., Singh, H., Spittal, M. J. (2016, January 28). Prevalence and Characteristics of Physicians Prone to Malpractice Claims. *New England Journal of Medicine, 374*, 354-362. http://dx.doi.org/10.1056/NEJMsa1506137 Retrieved May 3, 2017.

Vandevoort, L. G., Amrein-Beardsley, A. & Berliner, D. C. (2004, September 8). National board certified teachers and their students' achievement. *Education Policy Analysis Archives*, *12*(46). Retrieved October 4, 2004, from http://epaa.asu.edu/epaa/v12n46/.

Veltri, B. (2010). *Learning on other people's kids: Becoming a Teach For America teacher*. Charlotte, NC: Information Age Publishers.

About the Author

David C. Berliner
Arizona State University

David C. Berliner is Regents' Professor of Education, Emeritus, at Arizona State University. He has also taught at the Universities of Arizona and Massachusetts, at Teachers College and Stanford University, and at universities in Canada, Australia, The Netherlands, Denmark, Spain, and Switzerland. He is a member of the National Academy of Education, the International Academy of Education, and a past president of both the American Educational Research Association [AERA]

and the Division of Educational Psychology of the American Psychological Association [APA]. He has won numerous awards for his work on behalf of the education profession, and authored or co-authored over 400 articles, chapters and books. Among his best known works are the six editions of the text *Educational Psychology*, co-authored with N. L. Gage; *The Manufactured Crisis*, co-authored with B. J. Biddle; *Collateral Damage: How High Stakes Testing Corrupts American Education*, co-authored with Sharon Nichols; and *50 Myths and Lies That Threaten America's Public Schools, co-authored with Gene V Glass.* He co-edited the first *Handbook of Educational Psychology* and the books, *Talks to Teachers, Perspectives on Instructional Time*, and *Putting Research to Work in Your School.*

About the Guest Editors

Lorin W. Anderson
University of South Carolina (Emeritus)
anderson.lorinw@gmail.com
Lorin W. Anderson is a Carolina Distinguished Professor Emeritus at the University of South Carolina, where he served on the faculty from August, 1973, until his retirement in August, 2006. During his tenure at the University he taught graduate courses in research design, classroom assessment, curriculum studies, and teacher effectiveness. He received his Ph.D. in Measurement, Evaluation, and Statistical Analysis from the University of Chicago, where he was a student of Benjamin S. Bloom. He holds a master's degree from the University of Minnesota and a bachelor's degree from Macalester College. Professor Anderson has authored and/or edited 18 books and has had 40 journal articles published. His most recognized and impactful works are *Increasing Teacher Effectiveness, Second Edition*, published by UNESCO in 2004, and *A Taxonomy of Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, published by Pearson in 2001. He is a co-founder of the Center of Excellence for Preparing Teachers of Children of Poverty, which is celebrating its 14th anniversary this year. In addition, he has established a scholarship program for first-generation college students who plan to become teachers.

Maria de Ibarrola
Center for Research and Advanced Studies
mdeibarrola@gmail.com
Maria de Ibarrola is a Professor and high-ranking National Researcher in Mexico, where since 1977 she has been a faculty-member in the Department of Educational Research at the Center for Research and Advanced Studies. Her undergraduate training was in sociology at the National Autonomous University of Mexico, and she also holds a master's degree in sociology from the University of Montreal (Canada) and a doctorate from the Center for Research and Advanced Studies in Mexico. At the Center she leads a research program in the politics, institutions and actors that shape the relations between education and work; and with the agreement of her Center and the National Union of Educational Workers, for the years 1989-1998 she served as General Director of the Union's Foundation for the improvement of teachers' culture and training. Maria has served as President of the Mexican Council of Educational Research, and as an adviser to UNESCO and various regional and national bodies. She has published more than 50 research papers, 35 book chapters, and 20 books; and she is a Past-President of the International Academy of Education.

D. C. Phillips
Stanford University
d.c.phillips@gmail.com

D. C. Phillips was born, educated, and began his professional life in Australia; he holds a B.Sc., B.Ed., M. Ed., and Ph.D. from the University of Melbourne. After teaching in high schools and at Monash University, he moved to Stanford University in the USA in 1974, where for a period he served as Associate Dean and later as Interim Dean of the School of Education, and where he is currently Professor Emeritus of Education and Philosophy. He is a philosopher of education and of social science, and has taught courses and also has published widely on the philosophers of science Popper, Kuhn and Lakatos; on philosophical issues in educational research and in program evaluation; on John Dewey and William James; and on social and psychological constructivism. For several years at Stanford he directed the Evaluation Training Program, and he also chaired a national Task Force representing eleven prominent Schools of Education that had received Spencer Foundation grants to make innovations to their doctoral-level research training programs. He is a Fellow of the IAE, and a member of the U.S. National Academy of Education, and has been a Fellow at the Center for Advanced Study in the Behavioral Sciences. Among his most recent publications are the *Encyclopedia of Educational Theory and Philosophy* (Sage; editor) and *A Companion to John Dewey's "Democracy and Education"* (University of Chicago Press).

SPECIAL ISSUE
Historical and Contemporary Perspectives on Educational Evaluation

education policy analysis archives

Volume 26 Number 54 April 16, 2018 ISSN 1068-2341

Please send errata notes to Audrey Amrein-Beardsley at Audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.

archivos analíticos de políticas educativas
consejo editorial