

---

# Teacher Evaluation

Robert E. Stake

University of Illinois, Urbana-Champaign

Centre for Research in Applied Measurement and Evaluation  
University of Alberta, Edmonton, AB, Canada

Research Report No. CRAME 98-01

November, 1998

---

## TEACHER EVALUATION<sup>1</sup>

Robert Stake  
University of Illinois, Urbana-Champaign

God, are you with us tonight?

OF COURSE.

It would help me get started if we had a list of all the things that people do.

THERE, IT'S ON THE SCREEN.

I don't see it.

ADJUST THE FOCUS.

Oh. Would you rank them in order, the things people do most.

WHAT YOU ARE LOOKING FOR IS LOWER ON THE SCREEN.

Hmmm. 37th. Teaching is the 37th most common thing that people do. So that has to include informal as well as formal teaching. How common is evaluating teaching?

NINETEENTH.

And that has to mean both informal and informal evaluation of teaching. Well, God, as a matter of fact, how good is public school teaching today?

SON, YOU DON'T UNDERSTAND. EVALUATING IS A HUMAN CONSTRUCTION. I'M A RADICAL CONSTRUCTIVIST. I CREATED A THINKING, TEACHING, EVALUATING ADAM AND EVE. AND, AS I SAID AT THE TIME, IT WAS GOOD. FROM THEN ON, ALL EVALUATION HAS BEEN CREATED BY THE CREATURES OF THE EARTH.

So, the quality of teaching can't be more than the good and bad that people see. Even though we are sure there is lots below the surface that we can't get at, our evaluations have to be of what we experience. By "evaluation" I mean: coming to know the quality of the act, the merit and shortcoming of teaching.

We have been reminded that there's much informal teaching and informal evaluating, in the car, at work, and while schmoozing. We won't dwell on that. I will speak today primarily of formal teaching and formal evaluation.

In Europe they call professors teachers. So will I today. You may think it odd that I would speak of kindergarten and graduate school in the same breath. But, with so cathedral a beginning, I don't want to dwell upon mere nooks and crannies.

**The Unknown Quality of Teaching.** Those of us not in a classroom do not know the quality of teaching in that room. Those in the classroom know the teaching better, but still not well. None of us know the quality of teaching in a big collection of classrooms and schools. We don't know how good or how bad is the teaching on our campuses. We sometimes know some folks who like the teacher and some who don't, but we barely know anything about the quality of the actual teaching. You know much better how Barbara Budd does her job, how Bob Essensa does his, and how Bill Clinton does his job, than how your daughter's teachers have done theirs.

It is not that we can't get any agreement on good and bad. But what people agree on is pretty superficial. One of the pioneers in student rating of instructors, Dick Spencer<sup>2</sup>, was never happy with more than a single scale—because, he said, students rated quality of teaching only on whether or not they liked the instructor.

For most children in ordinary schools, who get almost continuous grading by teachers and periodic standardized testing, we have pretty good marks of student performance on a large selection of academic tasks. But the marks do little more than rank the students from high to low.

---

<sup>1</sup> A paper presented to a public audience at the University of Alberta, Edmonton, November 19, 1998. I have had assistance especially from Linda Mabry, Edith Cisneros, but countless others in the development of these ideas. Cisneros and I presented on "Evaluation of Teaching in Higher Education," AERA annual meeting, San Diego, April 17, 1998. Fine assistance was provided by Michael Ross and Michael Jodoin in conveying God's voice.

<sup>2</sup> Personal communication, long ago.

They don't tell how educated the children are becoming. According to Tom Romberg, they don't tell what they've been taught, how much they've been taught, and how well they've been taught (Romberg, 1995).

**The Complexity of Teaching.** We don't know quality of teaching partly because of the complexity of education even for an individual youngster. And difficult as it is for the individual, it is audacious to aggregate indicators of intellectual maturation across a diverse population living in an enormously intricate society. Our indicators are feeble. Dick Jaeger has concluded that we mislead ourselves about quality of education by supposing that quality of teaching is more or less equivalent to quality of student performance (Jaeger, 1992). What does student assessment tell us about quality of teaching?

Student test scores are only weakly related to quality of teaching. What determines level of student performance is some rich mix of genetic predisposition, infant nurturing, sibling rivalry, early childhood experience, peer interactivity, teen rebellion, exposure to language and word games, television, and schooling. Schools are good for children but schools cannot overcome deep deficits. For generations, widely believed in, schooling was almost the only force for raising children above what parents could provide. Now schooling is probably less influential than peer experiences or television. But even with television, cyberspace, and youth milieus, the schools continue to have unique contributions to make, varying with different philosophies and curricular emphases.

How well our school and university people are making that contribution is not reflected in such measures as the SAT and SAIP and TIMSS. Achievement tests have never been formally validated as indicators of teaching quality. Achievement tests have never been formally validated as indicators of provision of educational opportunity. In Chicago, I observe teachers of students with very poor scores who are, within their constraints, doing a very good job of providing opportunity for children to mature intellectually. Assessment has been at the heart of school functioning for a quarter of a century but it has not provided even a small idea of which teachers, which schools, and which cities are doing a good job of teaching.

I spent many days last spring in a Chicago classroom doing a case study of a sixth grade teacher, Kimberly Grogan. Kimberly is young, personable, dedicated. Kimberly makes mistakes. She does a splendid job of coaxing further interpretation out of Pedro, Angela, Harry, and all her pupils. Sometimes she creates a theatrical electricity, a moment here, a moment there. I have seen her crash a prime teaching situation to attend to a minor moral infraction. Other than to have you read my study, I know of no way to represent the quality of Kimberly's teaching.

**Inventory of responsibilities.** The purposes of education, aggregated across the profession, across researchers, the public and the primary beneficiaries, are far more complex than those represented in goal statements and formal assessments. Facts, theories, and reasoning are human needs, not just in isolation, but interactively, incrementally, in a range of contexts. We hold a vast inventory of expectations, beyond catalogue, partly ineffable, many becoming apparent only in disappointment as students fall short. That immense inventory is approximated by the informal assessments by teachers. What we expect teachers to be sensitive to and ameliorative of is immense. Can teacher evaluation reach so far?

As education in Canada and the US passed from teacher-driven recitation to government-driven accountability, the role of formal assessment magnified. The traditional quality-control of schooling, i.e., informal management, board oversight, parent compliance, state and province guidelines, and accreditation, have continued to be prominent in school operations. But with widespread perception that the quality of public education has fallen off, other means have been added to evaluate and to improve teaching and learning. In the last thirty years, formalized student testing has become the most used indicator of teaching quality.

In almost every school, at every grade level and in each subject matter, student achievement is being tested. The last decade has seen efforts to set standards for student achievement needed

to raise Canadian and American Education to leading world positions, resulting in some appearances of change but no clear grounds for knowing that our teaching is getting better or worse (Berliner & Biddle, 1994).

**Psychometric and pedagogic validation.** Standardized test development is probably the most technically sophisticated specialty in Education. Definitions and analytic procedures, at least at the major testing companies, are scrutinized, verified, codified and reworked. The traditional ethics of psychometrics call for extensive construct validation of the measurements to be used in schooling. And it is not enough that the instruments and operations be examined for accuracy, relevance and freedom from bias, but that independent measurements be used to confirm that scores indicate what we think they indicate. Sound test development is a slow and intricate procedure.

In the development of standardized student assessment instruments by the states and provinces, adequate validation has seldom taken place. Tests have been analyzed statistically to see that they are internally consistent but not that they mean what users think they mean. It is only a presumption that assessment scores indicate quality of teaching—although a commonplace claim in political and media dialogue. Proper validation research might tell us the strength or weakness of test-based conclusions about teaching quality. Those studies have not been commissioned. The validation of standardized testing as an indicator of teaching quality has not taken place.

As I have pointed out in a monograph on the validity of testing (Stake, 1995), there is a fundamental difference in psychometric and pedagogic perceptions of teaching and learning. Simply put, the psychometric perception is one of attained ability; the pedagogic is of attained experience. Our tests work hard to discern student fitness, a ranking of competence based on general learning ability, and on-demand readiness to perform common operations, and acquaintance with common information. It is not a subject-matter orientation, even though the tests are often labeled math or history achievement tests. They are ability tests.

The pedagogic perception of teaching and learning focuses on the presence of a learner in a teaching situation, sometimes passive and sometimes active, exposed to units of knowledge and skill, with obligation to perform a few things from immediate memory, but accumulating a vast experience, a savings account, from which the student can draw upon in future learning or work situation. Specific knowledge is important but relearning is more the emphasis than memory—at least as reflected in what teachers do.

Teachers do differ, of course, and some are more psychometrically oriented than others, that is, more highly concerned with establishing who are the better learners in the class. In a comprehensive evaluation of teaching, it is important to recognize these two orientations, one toward muscling up competitive fitness versus one toward nurturing personal experience.

**Halfway.** So, at halfway, the points I have made are:

- There is no "true" quality of teaching, it depends on what people see—and they differ.
- But there is such a thing as quality of teaching because, some of the time, many agree on what is good and bad teaching.
- Much of what they agree on is personalistic, pedagogically superficial, not taking into account what should have been taught and even whether or not it was taught.
- Education and teaching are far more complex than any testing, grading, or ratings used, emphasizing exposure and experience more than memory and routine.
- We know very little about how great or small the contribution to each student's education.

AREN'T YOU BEING A BIT PRECIOUS.

I beg your pardon.

DON'T THESE PEOPLE JUST WANT SOME ROUGH INDICATORS THEY CAN COUNT ON, LIKE BAD, POOR, FAIR AND GOOD?

I thought you were on my side.

PERHAPS I'M PLAYING DEVIL'S ADVOCATE.

I don't think you should want some simple rubrics for evaluating teaching. I think they hurt more than help, especially if they summarize quality in a single score.

Let me put it simply. Teachers are engaged in vital teaching functions, most of which are too complex and too unobservable for others, and sometimes even for themselves, to know.

Examples of these vital functions are:

- awareness and protection of teachable moments,
- discerning when a student has made a quantum leap,
- contending with students who would wrest away classroom control,
- honoring parent aspirations and fears, and
- linking new concepts with old.

We do not know how to evaluate these unseen functions. Still, a lot of what teachers do can be seen and it tells us whether or not they are inside the general territories of teacher responsibility. One can evaluate a teacher as poor for operating outside those territories, or good for staying within, or commendable for venturing into additional territories we admire. But these territorial evaluations are not glimpses, or correlates. They do not indicate the quality of the vital teaching functions.

**Supervisory evaluation.** Management of teaching cannot be effective without some assessment of teaching competence. The best and the worst we have is informal teacher evaluation, administrator-driven, sometimes capricious and sometimes more aimed at avoiding embarrassment than optimizing service to students. Administrative evaluation most often is unconscious review, that of a principal or department head intuitively valuing a teacher, but surfacing into formality when something goes wrong, when promotion or an increment of pay depends on it, or when "teacher of the year" is in the offing.

Intuition works surprisingly well. Teachers highly esteemed tend to be sensitive to what students are doing. Administrators overemphasize classroom management but many have a good idea which classrooms are enriching. Most students are in some ways enriched. But few incompetent teachers are dismissed, or helped get better, or even recognized. The teachers themselves have reservation about creative reviews but they tremble at the thought that a more rigorous system could be simplistic and punitive.

**Teacher Behavior**

Partial	.....	Fair
Autocratic	.....	Democratic
Aloof	.....	Responsive
Restricted	.....	Understanding
Harsh	.....	Kindly
Dull	.....	Stimulating
Stereotyped	.....	Original
Apathetic	.....	Alert
Unimpressive	.....	Attractive
Evading	.....	Responsible
Erratic	.....	Steady
Excitable	.....	Poised
Uncertain	.....	Confident
Inflexible	.....	Adaptable
Pessimistic	.....	Optimistic
Immature	.....	Integrated
Narrow	.....	Broad

When supervisors move into formal evaluating, they rely heavily on checklists and scales. Some do the occasional in-class observation. The personnel evaluation literature and organization development literature provide an abundance of criteria. In the open-ended sections of rating instruments, special merit is described, but seldom teacher impropriety, unless it has been decided that a case is to be made against the teacher.

The literature is replete with observation forms such as the classic bipolar list of attributes from Ryans (1960) shown above. Here a teacher can be seen as: "partial" or "fair," "autocratic" or "democratic," "aloof" or "responsive," "restricted" or "understanding," and so on. These are personality characteristics. But good teachers are found of any personality orientation, and too much of any kind in a school is problematic. Such traits do not take us very far toward sound teacher evaluation.

**Centering on Teacher Tasks.** What teachers do, in and out of class, needs to be at the center of the evaluation (Scriven, 1973; Borich, 1977). Accordingly, many school administrators have obtained forms ticking off regular duties: submits lesson plans on time; shares the chaperone load, is available to parents, inspires students to develop a portfolio. These are things many teachers should do but, in an evaluative scheme, these are pressures for compliance and teamwork as much as steps toward good teaching. I will mention again the need for diversity within any faculty; too much compliance is a bad thing, though what administrator would ever testify having had it.

College administrators get a special bump out of student ratings of instruction. Rating systems have been developed to an art, and need to be a part of any comprehensive teacher evaluation system, at least beginning at middle school. In ordinary classroom use, they cannot be treated as a representative sample of student population, even for one classroom, but as an important communication from the students who respond. One can learn as much from so called deviant responders as from the modal.

Whether at school or college, seldom do the evaluating supervisors delve into philosophy or foundational pedagogy. Years ago, when I had not seen anybody try it, I drafted a scale based on Harry Broudy's (1963) historic exemplars of teaching, the three purposes of teaching being: didactic, heuristic, and philetic.

**Correlates of good teaching.** I still have three points to make about formal evaluation of teaching, but I will move along quickly:

- No instrument or procedure should be used alone. If we don't have three or more ways of getting a look at teaching quality, we shouldn't use any. One view of the Indian elephant is not better than no view at all.
- A teacher should be evaluated on contributions to the entire instructional program, not just to his or her own classes.
- We can use existing research on teaching to suggest ways of improving teaching but we cannot use it for evaluating.

This last one is very complicated. Suppose the literature reports a strong correlation between providing accurate information as to when assignments are due and credible ratings of teaching quality. To help a teacher having difficulty with teaching, it's appropriate to look at information provided students. But to evaluate a teacher, we cannot use that correlational finding between assignment information and good teaching to mark a teacher good or bad.

Here's one that's clearer. Suppose a large research study concludes that teachers of a particular Canadian heritage, as a group, do poorer than average when teaching math classes. Should you evaluate any one particular teacher of this heritage as a poor teacher on those grounds? Of course not. It is ethically wrong, and partly on measurement grounds. In any real data correlation, there is some range of dependent variable scores for each level of input. Any one individual may be a person different from his or her subgroup.

When evaluating human beings, it's unethical to discriminate on any correlate of a criterion variable if the correlate was not included in public or contractual definition of good performance. Thus, unless they were required—we should not include in the evaluative criteria of effective teaching: whether or not the teacher repeats main points covered, or mumbling, intimidation, use of fair tests, humor, or knowledge of hockey. It is acceptable to evaluate on: literacy, sexist behavior, embezzlement and other things required or prohibited by law, or regulation, or job description. The point is that good research about pedagogical factors statistically associated with good teaching is not a sufficient ground for evaluating a particular teacher's teaching good or bad. Does that rule out some checklist items you thought were pretty good?

**Communitarian teaching.** And last or almost, I want to talk about communitarian teaching. Here the schools are more advanced than the colleges. The traditional concept of evaluating teaching is the evaluation of an autonomous instructor in an individual classroom.<sup>3</sup>

I'm urging an additional perspective, evaluating the contribution each instructor makes to the maintenance and improvement of all instructional programs in the department. What instructors do directly for students in their classes is, of course, important but what instructors contribute to the integrity of all offerings, not just their own, is important too. A charismatic lecturer or innovative lab-organizer or personalistic mentor, that is, a star, sometimes contributes little to the upgrade of weak, misdirected, frivolous and out-dated courses in the department. Both individual and team contributions need to be considered if teaching is to be evaluated at all.<sup>4</sup>

Collaboration across a campus faculty about matters of teaching is not new, but, in most places, it remains the exception more than the practice. Writing about a faculty as a "community of practice"<sup>5</sup> has become identified with Philip Morrison and John Seely Brown and colleagues at the Institute for Research on Learning. Some observers (Wenger, 1991; Brown, 1997; Alpert, 1998) of US and Canadian campuses have noted the scarcity of departments where instructors work closely together to maintain and improve teaching programs.<sup>6</sup> Wenger said:

*Even those who speak about learning organizations, life long learning, or the information society do so mostly in terms of individual learners and information processes. The notion of communities of practice helps us break this mold (p. 7).*

---

<sup>3</sup> A substantial body of research has been conducted along this line of inquiry. Most of the research has focused on methods and sources of information regarding teaching effectiveness, especially on the use of students as raters. (Kinney and Smith, 1992; Braskamp, Brandenburg & Ory, 1984; Cashin, 1988; Marsh, 1987; El-Hassan, 1995).

<sup>4</sup> Wheeler Loomis, once head of the Physics Department on our campus, used to keep a list of faculty names in his pocket, a list of those who contributed most to the department. According to Daniel Alpert (1998), Loomis recognized individual professional performance but also placed great value on qualities that held the community together. The Loomis List ranked individuals in terms of individual contribution to the department. At one time, at the top of the list was a Nobel Prize winner, so placed not for intellect but for his powerful contributions to other members of the department.

<sup>5</sup> A community of practice can be defined as "a group of professionals, informally bound to one another through exposure to a common class of problems, common pursuit of solutions, and thereby themselves embodying a store of knowledge." (Peter & Trudy Johnson-Lenz, 1998). Communities of practice have been traced back to the European guilds, but some writers see them as old as human interactivity of any kind (Community Intelligence Labs, 1998).

<sup>6</sup> According to Alpert (1998), "Learning is not only an activity, but also a vehicle for engagement with others. Learning is a social phenomenon. We all belong to communities of practice (work, school, in personal activities). It is through membership in communities of practice that we come to know—and become empowered by what we know. The social world is where work gets done, where learning takes place. Instructors encompass an ensemble of interconnected communities of practice whose boundaries do not necessarily (or usually) follow the formal boundaries of the organization"

**The Future.** Grounds for predicting future improvements in teaching seem to me to be tied to authentic assessment of student performance. That includes selection of testing tasks that more closely approximate real life behavior and challenge problem solving capabilities, tasks that require knowledge of complex topics and not just reasoning aptitudes (Berlak et al., 1992). But breakthrough depends on much more than better assessment tasks selected.

Any real advance probably will come, as Linda Mabry (in press) advocates, by teachers restructuring classroom assessment, using their value judgments not to register normative comparisons but to shape the maturation of the individual youngster.<sup>7</sup> This means more ipsative (person-referenced) assessment, less normative (norm group referenced) assessment. It is a vital step toward placing the well being of individual children or graduate students over national talent-sorting mechanisms. As I said earlier when I spoke of psychometric and pedagogical perspectives, the two policies are at odds. Teacher evaluation will sometimes need to support the teacher who nurtures experiences over the one who develops competitive fitness.

The school is stronger with a diverse staff, with teachers who serve a variety of role modelings, with different philosophies and cultural orientations. Differentiated staffing calls for different standards for individual faculty members. The quality of teaching then is not one rubric fits all but according to a different contract with each teacher.

**Closing.** God, I can't let these good people go home tonight without one answer. What is the quality of teaching in the Edmonton schools?

IT IS GOOD.  
 And could be better?  
 OF COURSE.  
 But how good is it?  
 I THINK YOU SHOULD READ ECCLESIASTES.  
 A season for all things?  
 NO, FURTHER ON.  
 Ah, you mean, this too is vanity!  
 THINK ABOUT IT.

So by my count, we are increasing the emphasis on evaluation without increasing support for professional development, for helping teachers improve. All too often, we don't know how to use good evaluative information when we get it. Under these circumstances, is it only vanity that pushes us further into assessing what a teacher is doing?

The critical question is not how to evaluate but for what will the results be used. To help or to take action against. With simple rubrics, I see the perils outnumbering the rewards. Teach to the test, now, teach to the rubric.

Perhaps the real hope for the future is in reducing appetites for measuring the quality of teaching, particularly pressures from the measurement and management communities. To decide that we cannot measure it sufficiently well and should not extend its imposition. We may come to recognize that to know really how effective and ineffective we are as teachers, undermines the good teaching we can do. Evaluating teaching well may depend on evaluating less.

---

<sup>7</sup> Both peer and adult judgments are to be respected.

## Bibliography

- Daniel A. Alpert, 1998. Lecture on the Representation of Quality. Seminar on Theories of Educational Evaluation, Spring, February 23. University of Illinois.
- Harold Berlak, Fred Newmann, E. Adams, Doug Archbald, T. G. Burgess, John Raven, and Tom Romberg, 1992. Toward a new science of educational testing and assessment. Albany: SUNY Press.
- David Berliner and Bruce Biddle, 1995. The manufactured crisis: Myths, fraud, and the attack on America's public schools. Reading, MA: Addison Wesley.
- Gary D. Borich, 1977. The appraisal of teaching. Reading, MA: Addison Wesley
- Larry Braskamp, Dale C. Brandenburg & John Ory, 1994 . Evaluating Teaching Effectiveness. London: Sage Publications.
- Harry S. Broudy, 1963. Historic exemplars of teaching method. In Nate L. Gage, editor, Handbook of Research on Teaching. Chicago: Rand McNally.
- John S. Brown, 1997. Common sense of purpose. In What is a community of practice. Community Intelligence Labs: <http://www.co-i-l.com/coil/knowledge-garden/cop/definitions.shtml>.
- Community Intelligence Labs, 1997. Communities of Practice. Internet address: <http://www.co-i-l.com/coil/knowledge-garden/cop/index.shtml>.
- W. E. Cashin, 1988. IDEA paper no. 20: Student ratings of teaching: A summary of the research. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.
- Cisneros-Cohernour, Edith. J. 1997. Trade-offs: The use of student ratings results and its possible impact on instructional improvement. University of Illinois: Unpublished report.
- Carmilva Flores, 1998. Extending teaching reform and professional development: The case of a school on probation. Unpublished doctoral dissertation, University of Illinois.
- K. El Hassan, 1995. Students' ratings of instruction: Generalizability of findings. Studies in Education, 21, pp. 411-429.
- Marit Granheim, Ulf Lundgren and Maurice Kogan, editors, 1990. Evaluation as policy making: Introducing evaluation into a national decentralized education system. London: Jessica Kingsley.
- Richard Jaeger, 1992. World class standards, choice, and privatization: Weak measurement serving presumptive policy. Vice presidential address, American Educational Research Association, San Francisco.
- Peter and Trudy Johnson-Letz, 1997. Bonding by exposure to common problems. In "What is a community of practice?" Community Intelligence Labs. Internet address: <http://www.co-i-l.com/coil/knowledge-garden/cop/definitions.shtml>.
- D. P. Kinney & S. P. Smith, 1992. Age and teaching performance. Journal of Higher Education, 63, 282-302.
- Linda Mabry, in press. Portfolios plus: A critical guide to alternative assessments and portfolios. Thousand Oaks, CA: Corwin Press.
- George Madaus, 1991. The effects of important tests on students: Implications for a national examination or system of examinations. AERA Conference on Accountability as a State Reform

Instrument: Impact on Teaching, Learning, Minority Issues and Incentives for Improvement. Washington, D.C.

H. W. Marsh, 1987. Student's evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253-388.

P. Morrison, 1986. Technology and culture, 8, 4, pp. 319-327.

Andrew Porter, 1989. External standards and good teaching. Educational Evaluation and Policy Analysis, 11, 4. 354.

H. H. Remmers, 1963. Rating Methods in Research on Teaching. In Nate L. Gage, editor, Handbook of Research on Teaching. American Educational Research Association. Chicago, IL: Rand McNally & Company.

Thomas A Romberg, editor, 1995. Reform in school mathematics and authentic assessment. Albany: SUNY Press.

D. G. Ryans, 1960. Characteristics of teachers. Washington, DC: American Council on Education.

Michael Scriven, 1974. The evaluation of teachers and teaching. California Journal of Education Research, 15, 3, 109-115.

E. Wenger, 1991. Communities of practice: Where learning happens. Benchmarks. Fall, pp. 6-8.