Should Student Outcomes Be Used to Evaluate Teaching?

Ronald A. Berk*
The Johns Hopkins University

A Context for Learning Outcomes

So what's the problem? Ever since the first brontosaurus stomped onto the screen in *Jurassic Park,* every source of evidence you could possibly use in the evaluation of teaching was FALLIBLE—from student ratings to peer observations to learning outcomes to ratings by close relatives. That applies to all *formative* (teaching and course improvement), *summative* (annual review, contract renewal, promotion & tenure), and *program* (accreditation and accountability) decisions. All of the sources are evil, chock full of psychometric sin. However, the different sources vary considerably in the type and degree of sin. They are like all the bad food we eat, except kale, which tastes like insulation unless you blend it into a smoothie with fruit, yogurt, flaxseed, and Doritos® to mask the flavor.

The most defensible strategy is to pick the best sources for a specific decision according to technical and legal standards. After all, high-stakes, career employment decisions are being made about faculty. The context for that strategy and use of outcome measures is briefly described in this section: (1) state of current practice, (2) 15 sources of evidence, and (3) triangulation of multiple sources.

State of Current Practice

Since the 1990s, give or take a decade, the practice of augmenting student ratings with other data sources of teaching effectiveness has been gaining traction in liberal arts colleges, universities, medical schools/colleges, and other institutions of higher education worldwide and in a few distant planets. Such sources can serve to *broaden and deepen the evidence base* used to evaluate courses and the quality of teaching (Arreola, 2007; Benton & Cashin, 2012; Berk, 2005, 2006, 2013a, 2013b; Cashin, 2003; Gravestock & Gregor-Greenleaf, 2008; Hoyt & Pallett, 1999; Knapper & Cranton, 2001; Seldin, 2006; Theall & Feldman, 2007). In fact, several comprehensive models of "faculty evaluation" have been proposed (Arreola, 2007; Berk, 2006, 2009a, 2009b; Braskamp & Ory, 1994; Centra, 1993; Gravestock & Gregor-Greenleaf, 2008), which include multiple sources of evidence.

15 Sources of Evidence

Guess what? There are 15 potential sources of evidence of teaching effectiveness reported in the literature: (1) student ratings, (2) peer observations, (3) peer review of course materials (4) external expert ratings, (5) self-ratings, (6) videos, (7) student interviews, (8) exit and alumni ratings, (9) employer ratings, (10) mentor's advice, (11) administrator ratings, (12) teaching scholarship, (13) teaching awards, (14) learning outcome measures, and (15) teaching (course) portfolio.

A critique and the major characteristics of each source, including type of measure needed to gather the evidence, the person(s) responsible for providing the evidence , the person or committee who uses the evidence, and the decision(s) typically rendered based on that data, were presented previously (Berk, 2006, 2013b). In fact, our hero's review should have been delivered to your doorstep by an Amazon drone. If you didn't get it, contact Amazon.

Triangulation of Multiple Sources

There are stacks of articles that weigh the merits and shortcomings of these various sources of evidence (Berk, 2005, 2006). Put simply: *There is no perfect source*

*The author is extremely grateful to *Steve Benton* (senior research officer & the only guy here with a real job, The IDEA Center), *Mike Theall* (professor emeritus, Youngstown State University), and *Bill Pallett* (former president, The IDEA Center) for reviewing the accuracy and appropriateness of the content and for their thoughtful technical comments on an earlier draft of this article. None of the reviewers should be held responsible for the recommendations and opinions expressed herein or any mistakes in the text. There is only one person who should be held accountable, and that person, of course, is: Peter Seldin.

or combination of sources, plus there is a scarcity of evidence on different combinations, such as student ratings and self-ratings (Barnett, Matthews, & Jackson, 2003; Stalmeijer et al., 2010). Each source can supply unique information, but, as noted previously, also is fallible, usually in ways different from the other sources. For example, peer ratings tend to be less reliable with biases that are different from student ratings (Thomas, Chie, Abraham, Raj, & Beh, 2014); student ratings have other psychometric weaknesses (Benton & Cashin, 2012; Nilson, 2012; Spooren, Brockx, & Mortelmans, 2013).

What should you do? *Draw on three or more different sources of evidence*. The strengths of each source can buffer the weaknesses of the other sources, thereby *converging on a decision about teaching effectiveness that is more accurate, reliable, equitable, and comprehensive than one based on any single source* (Appling, Naumann, & Berk, 2001). This notion of *triangulation* is derived from a compensatory model of decision making. It can be applied to teaching in a real-time, face-to-face class, hybrid-time class, an online virtual class, or a time-warp intergalactic class.

Why Use Learning Outcome Measures?

I'm sure you've asked yourself this question in a private moment, breakfast buffet, or raucous party. Learning outcome measures have received increased attention in recent years by state legislatures, government agencies, accreditation review boards at the state, regional, and national levels, public colleges and universities requiring evidence for accountability (Berrett, 2013), and by faculty needing more publications for promotion. The press for responsibility has been ratcheted up a few Emeril-notches. It's not a topic you can pluck from the headlines of your local newspaper (that's now online) yet, but that time is approaching.

In contrast, outcomes are rarely addressed in the teaching evaluation research (Clayson, 2009; Fenwick, 2001; Galbraith, Merrill, & Kline, 2012; Nilson, 2013; Seidel & Shavelson, 2007; Stark-Wroblewski, Ahlering, & Brill, 2007; Stehle, Spinath, & Kadmon, 2012). However, the topic has generated sufficient professional and public debate to emerge as one of the most contentious issues in education K–college. It would certainly qualify as a "flashpoint" (Berk, 2013b).

This section examines the conceptual underpinnings for applying student outcomes to instructors, but not students, as an: (1) indirect measure of teaching effectiveness based on the (2) factory worker–instructor productivity analogy.

Indirect Measure of Teaching Effectiveness

Among the aforementioned 15 sources of evidence, most involve *direct* ratings of teaching characteristics and behaviors. Learning outcome measures are a sticky and gooey source because they are *indirect*. Teaching performance is being inferred from students' performance—what they learned in the course. That relationship seems reasonable. After all, if you're an effective instructor, your students should perform well on measures of achievement and exhibit growth during the course in their knowledge of the subject. Of course, that assumes you don't have a class composed of students who—let's put it this way—when you look into their eyeballs, you can tell that the wheel is turning, but the hamster is out to lunch.

Despite the logic of using this source, only a paltry 5–7% of liberal arts colleges reported that they "always use" student exam performance or grade distribution for summative decisions related to teaching performance (Seldin, 1999). Those percentages have probably increased a smidgen in the past 15 years with the stiffening of accreditation requirements and state-wide accountability for public institutions.

Factory Worker–Instructor Productivity Analogy

If you are considering student outcomes, student achievement or growth is the measure of teaching effectiveness; that is, it is outcome based. If a factory worker's performance can be measured by the number of wickets (*Remember:* World Wide Wicket Company in "How to Succeed in Business without Really Trying!") he or she produces over a given period of time, why not evaluate an instructor's productivity by his or her students' success on outcome measures? (*NOTE*: This logic could be extended to other professions. For example, Aiken et al. (2014) studied the relationships between nurses' educational qualifications and patient and hospital outcomes. In medicine, measures of patients' health improvement over time, such as BMI, BP, girth, hue, and bodily fluid lab results, could be used to evaluate a physician's effectiveness.)

The arguments for this factory worker–instructor productivity analogy are derived from the principles of a piece-rate compensation system (Murnane & Cohen, 1986). Piece-rate contracts are the most common form of "payment by results" (Pencavel, 1977; Seiler, 1984). These contracts provide a strong incentive for workers to produce, because high productivity results in immediate rewards, possibly even a decent minimum wage and healthcare benefits.

When this "contract" concept is applied to teaching, it disintegrates for three reasons:

1. A factory worker uses the *same materials*, such as plywood and chewing gum, to make each wicket. Instructors work with students whose characteristics vary considerably within each class and from course to course.
2. The characteristics of a factory worker's *materials rarely influence his or her skills and rate of production*; that is, the quality and quantity of wicket production can be attributed solely to the worker. Instructors have no control over the individual differences and key characteristics of their *students*, such as ability, attitude, motivation, age, gender, ethnicity, cholesterol, and blood glucose, and of their *courses*, such as class size, composition, classroom facilities, available technology, and class climate. These characteristics can affect students' performance regardless of how well an instructor teaches.
3. The production of wickets is *easy to measure*. Just count them. Measuring students' performance on different outcomes is considerably more complicated with significant challenges to obtaining adequate degrees of reliability and validity for the scores. Then one has to pinpoint the component in the scores that is attributable to the instructor's teaching.

Consequently, the factory worker analogy just doesn't stick. It's like Teflon® to instructor evaluation. *Student outcomes provide a patina of credibility as a measure of teaching rather than an authentic source of evidence*. (*DERIVATION*: For you language scholars, *patina* is derived from two Latin words, *pa,* meaning literally "Opie's dad," and *tina-meana-bo-neena*, meaning "is Sheriff Andy Taylor.")

Critical Issues in the Use of Student Outcomes

After reading the preceding eight paragraphs, you may still be deliberating over the use of student outcomes as a source of evidence in your evaluation of instructors. Now let's complicate your deliberations by examining several critical issues that must be addressed: (1) selection of outcome measures, (2) measurement of achievement gain, (3) relationship between student ratings and learning outcomes, (4) isolating teaching effect with value-added models, and (5) technical and legal standards for personnel decisions.

Selection of Outcome Measures

What types of measures can be used to estimate achievement at one point in time, two points (pre- and posttesting), or multiple times (multiwave) during a course? The type, content, format, and psychometric properties of the measures can markedly affect the results.

Here are the options currently available:

1. Instructor-made measures (e.g., multiple-choice tests, essay tests, clinical exams, projects, problem-solving exercises, simulations)
2. Perceived learning measures
   - *Student Assessment of Their Learning Gains* (*SALG*) (http://www.salgsite.org)
   - *Transparency in Learning and Teaching Survey* (http://www.unlv.edu/provost/teachingandlearning)
   - *National Survey of Student Engagement* (*NSSE*) (http://www.nsse.iub.edu)
   - Knowledge surveys (Nuhfer, & Knipp, 2003; Wirth & Perkins, 2005)
3. Standardized tests
   - *ETS Proficiency Profile* (critical thinking, reading, writing, mathematics) (https://www.ets.org/proficiencyprofile/about/content)
   - Professional licensure and certification tests (e.g., teaching, accounting, nursing, medicine, taxidermy, skydiving, espionage)

Instructor-made measures are the easiest to administer and interpret for a single course. They are already being given as part of the course to assess students, so no additional instruments are needed. However, there are four problems or limitations: (1) they are typically the least reliable among the three categories of measures, (2) testing at one or more points in time is possible with parallel or equivalent forms for pre-post or multiple testings, but highly impractical, (3) the scores are content-specific, but not necessarily instructor-specific, and (4) the results are not generalizable to department, school, or institutional levels.

Perceived learning measures reviewed by Nilson (2013) do not measure actual learning. They are self-assessments by students of their learning experiences and different aspects of the course. Although there is some evidence that they are unrelated to actual learning (Weinberg, Fleisher, & Hashimoto, 2007; Weinberg, Hashimoto, & Fleisher, 2009), they do correlate highly with student achievement as do teachers' estimates of how much students have learned (Sudkamp, Kaiser, & Moller, 2012).

NSSE (particularly deep-learning scales) estimates how undergraduates spend their time and what they gain from attending college. It has been administered in more than 1500 colleges for institutional accountability to state legislatures. Course-specific knowledge surveys ask students to rate their level of confidence to answer

questions or perform tasks covering the course content and skills. Although these perceived learning measures provide inadequate proxies for achievement tests, they can furnish useful, but limited, information for evaluating instructors and for program evaluation.

Standardized tests are being administered in more than a dozen states in public institutions to document learning for accreditation and accountability (Berrett, 2013). The results are being compared across disciplines, colleges and universities, and state systems. Unfortunately, the content in these tests is usually unrelated to the curriculum and instruction in any single course, thereby rendering them inapplicable to instructor evaluation.

Measurement of Achievement Gain

Estimating pretest-posttest gain. Over the past 60 years, there have been mounds of research on how to measure change or gain over time (Bereiter, 1963; Chiou & Spreng, 1996; Cronbach & Furby, 1970; Linn, 1981; Linn & Slinde, 1977; Lord, 1956; O'Connor, 1972; Smolkowski, 2010; Zimmerman & Williams, 1982, 1998). A variety of methods have been proposed for estimating gain over two measurement points (pretest and posttest), including raw gain, gain adjusted for pretest error, gain adjusted for pretest and posttest error, the difference between true posttest and pretest scores, raw residual gain, estimated true residual gain, a "base-free" procedure, and posttest score adjusted for initial academic potential.

Deficiencies of gain scores. Two major deficiencies of pretest-posttest gain scores have been cited in the literature over and over again: (1) their low reliability and (2) their negative correlation with pretest scores. The findings of investigations comparing the numerous strategies for estimating gain (Nesselroade, Stigler, & Baltes, 1980; Overall & Woodward, 1975, 1976; Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983, 1985) concluded that those deficiencies are not serious; they are misconceptions rather than deficiencies. Not to worry.

The unreliability of gain scores should not be a cause for concern in determining an instructional effect between two testings. The negative bias of the correlation should be interpreted as an artifact of measurement error on the estimation of the correlation. However, there are other statistical and design artifacts, such as testing and regression effects, that can spuriously inflate or deflate the estimates depending on the method used to compute gains.

Multiwave data. Despite all of the buckets filled with mounds of research on the technical problems with gain scores, probably the most significant flaw is the meager information they yield based on only two measurement points. The use of *multiwave data,* where three or more measurements are obtained, vastly improves the measurement of change simply because additional information on all students is incorporated (Rogosa et aI., 1982) and it provides greater precision in estimating gain (Bryk & Raudenbush, 1987; Rogosa & Willett, 1985; Willett, 1988). The only side effect from multiwave data is a compulsion to eat cucumbers, which is certainly better than kale.

Relationship between Student Ratings and Learning Outcomes

Convergent validity. If learning outcomes are to be considered as a sidekick for student ratings and peer observations to improve teaching, what do outcome measures contribute to the evaluation process? Do they measure the same construct as student ratings or something different?

If the evaluation process involves multiple measures, student ratings should be the anchor. "Why?" you query. Excellent single-word question. They have a research base that spans 90 years, plus they yield multiple observations by multiple raters (Benton & Cashin, 2012, Berk, 2013b). As the reigning primary source of data on teaching throughout higher education for more than half a century, student ratings are a necessary, but not sufficient, source to evaluate teaching effectiveness (Berk, 2006, 2013b).

How do students' ratings of their instructors relate to their learning as defined by course tests, grades, student perceptions of learning, and standardized tests? Those relationships can furnish convergent validity evidence on the construct being measured. Let's examine the research.

Meta-analyses. More than 80 correlational studies have examined the relationship between student ratings and achievement based on common examinations given in courses with several sections. Meta-analyses of this research by Cohen (1981), d'Apollonia and Abrami (1997a, 1997b), Feldman (1989), and Onwuegbuzie, Daniel, and Collins (2009) aggregated the results to produce significant mean correlations ranging from .10 to .47. Another meta-analysis by Clayson (2009) found an average correlation of .13 between student ratings and test results. (*REMEMBER:* Coefficients below .50 denote that achievement explains less than 25% of the variance of student ratings; 75% are explained by other factors.)

Other validity studies. Braun and Leidner (2009) and Stapleton and Murkison (2001) reported correlations ranging from .28 to .75 between students' self- reported acquisition of competence (perceptions of learning) and their satisfaction with teaching (student ratings). Gal-

braith et al. (2012) found a nonlinear association between student ratings and a standardized test, while Stark-Wroblewski et al. (2007) found little or no relationship between student ratings and student learning. Stehle et al. (2012) reported similar results between student ratings and the subjective perception of learning with a multiple-choice test, but significant correlations with a practical clinical examination.

Conclusions. Overall, there seems to be no consensus on the degree of association between student ratings and learning outcome measures (Spooren et al., 2013).There was considerable variation in the validity coefficients reported in the preceding studies, not just in size, but in direction as well. This enormous amount of wiggle room in the coefficients suggests that there are many other factors that account for the learning. It tempers the notion that students give high ratings to instructors from whom they learn the most and low ratings to instructors from whom they learn the least. The low to moderate magnitude of the correlations indicate the convergent validity evidence is putrid.

Wiggle-wise, possible explanations for these results include: (1) the quality of the scales and outcome measures, (2) the variability of the outcome measures used in the studies (multiple-choice tests, grades, perceptions of learning, practical examination) (Stehle et al., 2012), (3) the instructors' characteristics, and (4) the students' knowledge of their final grade before rating the instructor.

Isolating Teaching Effectiveness with Value-Added Models

How to isolate teaching. Using students' performance on learning outcomes as an independent measure of teaching effectiveness is fraught with numerous difficulties. Berliner (2005) cautioned against this approach. The crux of the problem is this: *How do you isolate teaching as the sole explanation for student learning?* Performance throughout a course on tests, projects, reports, parties, chili cook-offs, and other indicators may be influenced by the characteristics of the students and course as well as the outcome measures themselves, over which faculty have no control (Berk, 1988, 1990).

Value-added models (VAMs). Crux-wise, VAMs of faculty evaluation focus on estimating an instructor's contribution to students' achievement and pre-post gains after statistically removing extraneous variables. Those variables can include characteristics of the

1. *students* (ability, gender, race/ethnicity, age, socioeconomic background, educational preparation, prior knowledge, interest, motivation, attitude, effort, attendance),
2. *courses* (size, attendance rate, heterogeneity, discipline, workload, climate, classroom facilities, available technology and learning resources, difficulty, level, type, elective vs. required), and
3. *outcome measures* (objective test, essay, performance test, project, perceived learning, standardized test, content, difficulty, discrimination, reliability, and content, criterion-related, construct, instructional, and curricular validity) (Murphy, Hallinger, & Heck, 2013).

To these 40 characteristics can be added testing and regression effects based on the pre-post design. The VAMs have received significant attention in the evaluation of K–12 teachers (Berk, 1988; Danielson, 2007; Gates Foundation, 2013; Hanushek & Rivkin, 2010; Hill, Kaptula, & Umland, 2011) and are now creeping into higher education.

Conclusions. Murphy et al. (2013) concluded that the evidence for using VAMs to evaluate teaching is insufficient due to the inconsistent and overstated magnitude of the effects of teaching-related variables on student learning outcomes and gains (Baker et al., 2010; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Rothstein, 2009). In other words, there is *an intractable problem isolating experimentally or statistically what an individual instructor contributes to student learning* beyond all of the other 42 possible student, course, outcome, and design variables mentioned previously. There is no known method that can furnish the surgical precision required to extract the teaching component. Moreover, this same verdict was reached 25 years ago in the application of student outcomes to K–12 teaching evaluation (Berk, 1988), a kind of déjà vudict.

How does one identify the optimal mix of instructional strategies, behaviors, and skills that account for student learning? That mix may be lurking in studies that have yet to be designed. There are also serious concerns about the validity, accuracy, and equity of value-added scores assigned to individual instructors (Hill et al., 2011).

Technical and Legal Standards for Personnel Decisions

Technical standards. In order for any source of evidence to be used for personnel decisions about faculty, it must satisfy standards for psychometric quality, especially *reliability and validity related to teaching effectiveness*. Explicit standards are given in the *Standards for Educational and Psychological Testing* (*SEPT*) (AERA, APA, & NCME Joint Committee on Standards, 1999) and the *Personnel Evaluation Standards* (*PES*) (Joint Commit-

tee on Standards for Educational Evaluation, 2009). The *SEPT* and *PES* standards specify that the measure must exhibit a *close link between the job content and behaviors and the items* (Standards 14.8 & 14.9). The specific reliability and validity standards were described previously (Berk, 2013b). Student outcomes do not satisfy any of those standards, nor are they mentioned as a possible source of evidence.

If student achievement is reported at the program, department, school, or institutional level for program assessment requirements in the context of accreditation and accountability (Berrett, 2013), the tests must also meet similar technical standards described in the *Program Evaluation Standards* (Yarbrough, Shulha, Hopson, & Caruthers, 2011). Since the inference is about program effectiveness, not individual students or instructors, the perceived learning measures and standardized tests may provide appropriate information. Instructor-made tests would not satisfy the standards.

Legal requirements. Beyond these technical standards, there are federal laws and court cases that designate legal requirements for employee decisions. Since instruments can be used to hire, promote, demote, or fire an employee, the U.S. Equal Employment Opportunity Commission's (EEOC) *Uniform Guidelines on Employee Selection Procedures* (U.S. Code of Federal Regulations, 1978) set forth laws to protect an innocent employee from an evil employer who intentionally uses them to discriminate based on pay, age, color, disability, national origin, pregnancy, race, religion, and sex (U.S. EEOC, 2010). The guidelines to measure job performance related to these federal anti-discrimination laws were described elsewhere (Berk, 2013b), as were the court cases on employment testing practices (Ashe & U.S. EEOC, 2007; Nathan & Cascio, 1986; Wines & Lau, 2006).

Conclusions. Based on the research and standards reviewed in the preceding sections, student learning outcomes defined as instructor-made measures (unknown or inadequate score reliability and validity), perceived learning measures (unrelated to actual learning and teaching behaviors), and standardized achievement tests (unrelated to course curriculum and instruction) lack sufficient validity and reliability evidence as legitimate measures of an instructor's teaching. *There is no link between any of these forms of student outcomes and teaching job behaviors and content*, a deal-breaker for use in personnel decisions. That flaw is fatal to the use of student outcomes alone or in conjunction with other measures. If student assessment standards are met, the measures can be used for program decisions of teaching methods and program improvement for accreditation and accountability at the institutional level.

Recommendations

When multiple sources of evidence are used to measure teaching effectiveness, the role of each source should be clearly defined before combining the different sources. The decision maker should *integrate the information from only those sources for which the highest levels of reliability* (Standards 2.1 & 2.7) *and validity* (Standard 14.13) *evidence are available*.

Start with the specific decision you need to make. That decision will drive your choice of sources. Then pick the most appropriate and technically sound sources. Now it's time to deliver the verdicts that you have anticipated for at least five paragraphs. For what decisions can learning outcomes be used?

Formative Decisions

Instructors in face-to-face, online, blended/hybrid, and intergalactic courses can use the student assessment tools they already developed to measure their students' performance. Here's how those scores can be leveraged for feedback on their teaching:

1. *Multiple measures*. Whether the information is scores at one point in time, gain scores, or multiwave scores, additional evidence from other measures of teaching, such as student ratings, peer observations, and self-ratings, should be collected to guide teaching and course improvements.
2. *Pre-posttest gain scores*. An instructor can administer a test at the beginning and end of a unit or course and use the difference scores as viable estimates of student growth. Computing gain scores between any two points would require the instructor to either (a) administer the same measure twice, such as pre- and posttesting, or (b) systematically develop a content and technically equivalent form (classically or randomly parallel form) to compare with the original measure. The former approach can be confounded by "testing and regression effects," which can artificially inflate or deflate the students' posttest scores; the latter is impractical in most any classroom setting.
3. *Multiwave data.* Although gain scores can provide limited feedback on teaching, it is still preferable to use multiwave data. They furnish a better estimate of true growth than the two snapshots used for simple gain scores. Further, an instructor can generate trajectory plots that summarize student growth over time. In other words, an instructor can use the scores from multiple measures at different time points over the semester to gauge the students' progress longitudinally as well as to infer his or her teaching effectiveness to some degree.

4. *Inferences for teaching improvement.* Extreme caution (*Code*: Tangerine) should be observed in interpreting test or gain scores alone as evidence for teaching improvement. Inferences should be drawn in conjunction with the direct data sources described previously.

Summative Decisions

Employment decisions about full-time faculty can involve annual review and feedback, contract renewal, merit pay, teaching awards, and promotion and tenure. Part-time or adjunct instructors are usually evaluated for contract or really-brief letter renewal. The sources of evidence used for these decisions must satisfy the standards cited in the preceding section.

Surprise! Surprise! There are two major, somewhat draconian conclusions:

1. *Technical and legal standards. These standards disqualify learning outcomes in any form* (instructor-made measures, perceived learning measures, or standardized tests) *as a reliable, valid, and equitable source* by itself or in conjunction with other sources.
2. *Gain scores*. Students' gain scores can be computed on any outcome measure. The issue is the inferences drawn from those scores. They may not be attributable solely, or, in some cases, even mostly, to the effectiveness of the teaching—*students' performance ≠ teaching performance*.

   There is no Value-Added Model for a pre-posttest design (with the same measure or parallel test forms) that can isolate the teaching element in student gains. In other words, there is no way to statistically cauterize the 42 confounding variables in order to cleanly extract the teaching effect from the total gain. This is tantamount to a "guilty verdict" in a murder case on *Law & Order* (Ka Chung!).

Program Decisions

All measures of learning outcomes can be used to satisfy accreditation and accountability requirements if they meet the technical standards for student assessment. Standardized student achievement test scores and perceived learning scores at the course level can be aggregated at the department, school, and institutional levels. Instructor-made measures could only furnish course-by-course evidence of students' performance. These measures of students' achievement and opinion about their learning can be reported to address the documentation needs in specific sections of an accreditation self-study. Teaching effectiveness should not be inferred from the aggregated scores or gain scores computed within courses.

References

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education) Joint Committee on Standards. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Aiken, L., H., et al. (2014). Nurse staffing and education and hospital mortality in nine European countries: A retrospective observational study. *The Lancet*. Retrieved on February 26, 2014, from http://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2813%2962631-8/abstract. DOI:10.1016/S0140-6736(13)62631-8

Appling, S. E., Naumann, P. L., & Berk, R. A. (2001). Using a faculty evaluation triad to achieve evidenced-based teaching. *Nursing and Health Care Perspectives, 22*, 247–251.

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building, and operating large-scale faculty evaluation systems* (3rd ed.). San Francisco: Jossey-Bass.

Ashe, L., & U.S. EEOC. (2007, May). Employment testing and screening: Recent developments in scored test case law. Retrieved on February 6, 2014, from http://eeoc.gov/eeoc/meetings/archive/5-16-07/testcase_ashe.html.

Baker, E. L., Barton, P. E., Darling-Hammond, L, Haertel, E, Ladd, H. F., Linn, R. L., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper 278). Washington, DC: Economic Policy Institute.

Barnett, C. W., Matthews, H. W., & Jackson, R. A. (2003). A comparison between student ratings and faculty self-ratings of instructional effectiveness. *Journal of Pharmaceutical Education, 67*(4), Article 117.

Benton, S. L., & Cashin, W.E. (2012). *Student ratings of teaching: A summary of research and literature* (IDEA Paper No. 50). Manhattan, KS: The IDEA Center. Retrieved on February 11, 2014, from http://www.theideacenter.org/sites/default/files/idea-paper_50.pdf.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison: University of Wisconsin Press.

Berk, R. A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education, 1*, 345–363.

Berk, R. A. (1990). Limitations of using student achievement data for career ladder promotions and merit pay decisions. In J. V. Mitchell, Jr., S. L. Wise, & B. S. Plake (Eds.), *Assessment of teaching: Purposes, practices, and implications for the profession* (pp. 261–306). Hillsdale, NJ: Erlbaum. Retrieved on February 2, 2014, from http://digitalcommons.unl.edu/burosassessteaching/10/

Berk, R. A. (2005). Survey of 12 strategies for measuring teaching effectiveness. *International Journal on Teaching and Learning in Higher Education, 17*(1), 48–62. Retrieved on January 28, 2014, from http://www.isetl.org/ijtlhe/pdf/IJTLHE8.pdf

Berk , R. A. (2006). *Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment, and decision making for faculty, administrators, and clinicians.* Sterling, VA: Stylus Publishing.

Berk, R. A. (2009a). Beyond student ratings: "A whole new world, a new fantastic point of view." *Essays on Teaching Excellence, 20*(1). (http://www.podnetwork.org/publications/teachingexcellence/05-06/V17,%20N2%20Berk.pdf)

Berk, R. A. (2009b). Using the 360° multisource feedback model to evaluate teaching and professionalism. *Medical Teacher, 31*(12), 1073–1080. DOI: 10.3109/01421590802572775

Berk, R. A. (2013a). Top 5 flashpoints in the assessment of teaching effectiveness. *Medical Teacher, 35*, 15–26. DOI: 10.3109/0142159X.2012.732247 (http://informahealthcare.com/doi/abs/10.3109/0142159X.2012.732247)

Berk, R. A. (2013b). *Top 10 flashpoints in student ratings and the evaluation of teaching: What faculty and administrators must know to protect themselves in employment decisions*. Sterling, VA: Stylus Publishing.

Berliner, D. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, *56*(3), 205–213.

Berrett, D. (2013, November 1). States demand that colleges show how well their students learn. *The Chronicle of Higher Education, 59* (9), A6. Retrieved on January 16, 2014, from http://chronicle.texterity.com/chronicle/20131101a?sub_id=WqzwGDqWviFn#pg6

Braskamp, L. A, & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Braun, E., & Leidner, B. (2009). Academic course evaluation: Theoretical and empirical distinctions between self-rated gain in competencies and satisfaction with teaching behavior. *European Psychologist, 14*, 297–306. DOI: 10.1027/1016-9040.14.4.297

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–159.

Cashin, W. E. (2003). Evaluating college and university teaching: Reflections of a practitioner. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 531–593). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco: Jossey-Bass.

Chiou, J., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, 9,* 158–167.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16–30. Retrieved on February 2, 2014, from http://jmd.sagepub.com/content/31/1/16.full.pdf+html

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281–309.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" or should we? *Psychological Bulletin, 74*, 68–80.

d'Apollonia, S., & Abrami, P. C. (1997a). Scaling the ivory tower, Part 1: Collecting evidence of instructor effectiveness. *Psychology Teaching Review, 6,* 46–59.

d'Apollonia, S., & Abrami, P. C. (1997b). Scaling the ivory tower, Part 2: Student ratings of instruction in North America. *Psychology Teaching Review, 6,* 60–76.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30,* 583-–645.

Fenwick, T. J. (2001). Using student outcomes to evaluate teaching. A cautious exploration. In C. Knapper & P. Cranton (Eds.), *Fresh approaches to the evaluation of teaching* (New Directions for Teaching and Learning, No. 88, pp. 63–74). San Francisco: Jossey-Bass.

Galbraith, C., Merrill, G., & Kline, D. (2012). Are student evaluations of teaching effectiveness valid for measuring student outcomes in business related classes? A neural network and Bayesian analysis. *Research in Higher Education, 53*, 353–374. DOI: 10.1007/s11162-011-9229-0

Gates Foundation. (2013). *Ensuring the fair and reliable measures of effective teaching (MET): Culminating findings from the MET Project's three-year study*. Seattle, WA: Author. Retrieved on February 6, 2014, from http://www.gatesfoundation.org/united-states/Pages/measures-of-effective-teaching-fact-sheet.aspx.

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends.* Toronto: Higher Education Quality Council of Ontario. E-book retrieved on February 2, 2014, from http://www.heqco.ca/en-A/Research/Research%20Publications/Pages/Home.aspx.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*, 267–271.

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

Hoyt, D. P., & Pallett, W. H. (1999). *Appraising teaching effectiveness: Beyond student ratings* (IDEA Paper No. 36). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.

Joint Committee on Standards for Educational Evaluation. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Knapper, C., & Cranton, P. (Eds.). (2001). *Fresh approaches to the evaluation of teaching* (New Directions for Teaching and Learning, No. 88). San Francisco: Jossey-Bass.

Linn, R. L. (1981). Measuring pretest-posttest performance changes. In R. A. Berk (Ed.), *Educational evaluation methodology: The state of the art* (pp. 84–109). Baltimore: Johns Hopkins University Press.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre and posttesting periods. *Review of Educational Research, 47,* 121–150.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16,* 421–437.

McCaffrey, D. L., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4,* 572–606.

Murnane, R. J., & Cohen, D. K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review, 56*, 1–17.

Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation: The case of the missing clothes. *Educational Researcher, 42*(6), 349–354.

Nathan, B. R., & Cascio, W. F. (1986). Introduction. Technical and legal standards. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 1–50). Baltimore, MD: Johns Hopkins University Press.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88*, 622–637.

Nilson, L. B. (2012). Time to raise questions about student ratings. In J. E. Groccia & L. Cruz (Eds.), *To improve the academy: 31. Resources for faculty, instructional, and organizational development* (pp. 213–228). San Francisco: Jossey-Bass.

Nilson, L. B. (2013). Measuring student learning to document teaching effectiveness. In J. E. Groccia & L. Cruz (Eds.), *To improve the academy: 32. Resources for faculty, instructional, and organizational development* (pp. 287–300). San Francisco: Jossey-Bass.

Nuhfer, E. B., & Knipp, D. (2003). The knowledge survey: A tool for all reasons. In C. Wehlburg & S. Chadwick-Blossey (Eds.), *To improve the academy: 21. Resources for faculty, instructional, and organizational development* (pp. 59–78). Boston, MA: Anker.

O'Connor, E. F. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research, 42*, 73–98.

Onwuegbuzie, A. J., Daniel, L. J., & Collins, K. M. T. (2009). A meta-validation model for assessing the score validity of student teaching evaluations. *Quality & Quantity, 43,* 197–209. DOI: 10.1007/s11135-007-9117-4

Overall, J . E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82,* 85–86.

Overall, J. E., & Woodward, J. A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin, 83*, 776–777.

Pencavel, J. H. (1977). Work effort, on-the-job screening, and alternative methods of remuneration . In R. Ehrenberg (Ed.), *Research in labor economics* (Vol. 1, pp. 225–258). Greenwich, CT: JAI Press.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726–748.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement, 20,* 335–343.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50,* 203–228.

Rothstein, J. (2009, January). *Student sorting and bias in value added estimation: Selection on observables and nonobservables* (NBER Working Paper, 14666). Cambridge, MA: National Bureau of Economic Research. Retrieved on February 6, 2014, from http://www.nber.org/papers/w14666

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis research. *Review of Educational Research, 77*, 454–499.

Seiler, E. (1984). Piece rate vs. time rate: The effect of incentives on earnings. *Review of Economics and Statistics, 66*, 363–375.

Seldin, P. (1999). Current practices – good and bad – nationally. In P. Seldin & Associates (Eds.), C*hanging practices in evaluating teaching: A practical guide to improved faculty performance and promotion/ tenure decisions* (pp. 1–24). Bolton, MA: Anker.

Seldin, P. (2006). Building a successful evaluation program. In P. Seldin & Associates (Eds.), *Evaluating faculty performance: A practical guide to assessing teaching, research, and service* (pp. 1–19). Bolton, MA: Anker.

Smolkowski, K. (2010, April). Gain score analysis. Retrieved on February 3, 2014, from http://homes.ori.org/~keiths/Files/Tips/Stats_GainScores.html.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598–642. DOI: 10.3102/0034654313296870

Stalmeijer, R. E., Dolmans, D. H., Wolfhagen, I. H., Peters, W. G., van Coppenolle, L., & Scherpbier, A. J. (2010). Combined student ratings and self-assessment provide useful feedback for clinical teachers. *Advances in Health Science Education, Theory, and Practice, 15*(3), 315–328.

Stapleton, R. J., & Murkison, J. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education, 25,* 269–291. DOI: 10.1177/105256290102500302

Stark-Wroblewski, K., Ahlering, R. F., & Brill, F. M. (2007). Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of teaching with pre-post learning measures. *Assessment & Evaluation in Higher Education, 32*, 403–415. DOI: 10.1080/02602930600898536

Stehle, S., Spinath, B., & Kadmon, M. (2012). Measuring teaching effectiveness: Correspondence between students' evaluations of teaching and different measures of student learning. *Research in Higher Education*. DOI: 10.1007/s11162-012-9260-9

Sudkamp, A., Kaiser, J., & Moller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743–762.

Theall, M., & Feldman, K. A. (2007). Commentary and update on Feldman's (1997) "Identifying exemplary teachers and teaching: Evidence from student ratings." In R. P. Perry & J. C. Smart (Eds.), *The teaching and learning in higher education: An evidence-based perspective* (pp. 130–143). Dordrecht, The Netherlands: Springer.

Thomas, S., Chie, Q. T., Abraham, M., Raj, S. J., & Beh, L-S. (2014). A qualitative review of literature on peer review of teaching in higher education: An application of the SWOT framework. *Review of Educational Research, 84*(1), 112–159. DOI: 10.3102/0034654314996173

U.S. Code of Federal Regulations. (1978, August 25). *Uniform guidelines on employee selection procedures.* 29 CFR part 1607, section 6A. Retrieved on February 10, 2014, from http://www.gpo.gov/fdsys/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml.

U.S. Equal Employment Opportunity Commission (EEOC). (2010, September). Employment tests and selection procedures. Retrieved on February 10, 2014, from http://www.eeoc.gov/policy/docs/factemployment_procedures.html.

Weinberg, B. A., Fleisher, B. M., & Hashimoto, M. (2007). Evaluating methods of evaluating instruction: The case of higher education. (NBER Working Paper No. 12844.) Retrieved on February 5, 2014, from http://www.nber.org/papers/w12844

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *Journal of Economic Education, 40*(3), 227-261. Retrieved on February 5, 2014, from http://dx.doi.org/10.3200/JECE.40.3.227-261

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–422). Washington, DC: American Educational Research Association.

Wirth, K. R., & Perkins, D. (2005, April 2). *Knowledge surveys: An indispensable course design and assessment tool*. In Proceedings of the Innovations in the Scholarship of Teaching and Learning Conference, Northfield, MN. Retrieved on February 12, 2014, from http://www.macalester.edu/geology/wirth/WirthPerkinsKS.pdf

Wines, W. A., & Lau, T. J. (2006). Observations on the folly of using student evaluations of college teaching for faculty evaluation, pay, and retention decision and its implications for academic freedom. *William & Mary Journal of Women and the Law*, *13*(1), 167–202. (http://works.bepress.com/cgi/viewcontent.cgi?article=1007&context=terence_lau)

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*, 149–154.

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology, 51*, 343–351.

Ronald A. Berk, Ph.D., is professor emeritus, biostatistics and measurement, and former assistant dean for teaching at The Johns Hopkins University. Now he is a full-time speaker, writer, PowerPoint coach, and jester-in-residence. He can be contacted at rberk1@jhu.edu, www.ronberk.com, www.pptdoctor.net, or www.linkedin.com/in/ronberk/, and blogs at http://ronberk.blogspot.com.