Top five flashpoints in the assessment of teaching effectiveness

RONALD A. BERK
Johns Hopkins University, USA

Abstract

Background: Despite thousands of publications over the past 90 years on the assessment of teaching effectiveness, there is still confusion, misunderstanding, and hand-to-hand combat on several topics that seem to pop up over and over again on listservs, blogs, articles, books, and medical education/teaching conference programs. If you are measuring teaching performance in face-to-face, blended/hybrid, or online courses, then you are probably struggling with one or more of these topics or flashpoints.

Aim: To decrease the popping and struggling by providing a state-of-the-art update of research and practices and a "consumer's guide to trouble-shooting these flashpoints."

Methods: Five flashpoints are defined, the salient issues and research described, and, finally, specific, concrete recommendations for moving forward are proffered. Those flashpoints are: (1) student ratings vs. multiple sources of evidence; (2) sources of evidence vs. decisions: which come first?' (3) quality of "home-grown" rating scales vs. commercially-developed scales; (4) paper-and-pencil vs. online scale administration; and (5) standardized vs. unstandardized online scale administrations. The first three relate to the sources of evidence chosen and the last two pertain to online administration issues.

Results: Many medical schools/colleges and higher education in general fall far short of their potential and the available technology to comprehensively assess teaching effectiveness. Specific recommendations were given to improve the quality and variety of the sources of evidence used for formative and summative decisions and their administration procedures.

Conclusions: Multiple sources of evidence collected through online administration, when possible, can furnish a solid foundation from which to infer teaching effectiveness and contribute to fair and equitable decisions about faculty contract renewal, merit pay, and promotion and tenure.

Introduction

> *FLASHPOINT: a critical stage in a process, trouble spot, discordant topic, or lowest temperature at which a flammable liquid will give off enough vapor to ignite.*

If you have read any of my previous articles, you know I have given off buckets of vapor. For you language scholars, "flashpoint" is derived from two Latin words, "flashus," meaning "your shorts," and "pointum," meaning, "are on fire."

Why flashpoints?

This article is not another review of the research on student ratings. It is a *state-of-the-art update of research and practices, primarily since 2006* (Berk 2006; Seldin & Associates 2006; Arreola 2007), *with specific TARGETS:* the flashpoints that have emerged, which are critical issues, conflicts, contentious problems, and volatile hot buttons in the assessment of teaching effectiveness. They are the most prickly, thorny, vexing, and knotty topics that every medical school/college and institution in higher education must confront.

These flashpoints cause confusion, misunderstanding, dissension, hand-to-hand combat, and, ultimately, inaccurate and unfair decisions about faculty. Although there are many more than five in this percolating cauldron of controversy, the ones tackled here seem to pop up over and over again on listservs, blogs, articles, books, and medical education/teaching conference programs, plus they generate a firestorm of debate by

Practice points

- Polish your student rating scale, but start building multiple sources of evidence to assess teaching effectiveness.
- Match your highest quality sources to the specific formative and summative decisions using the 360° MSF model.
- Review current measures of teaching effectiveness with your faculty and plan specifically how you can improve their psychometric quality.
- Design an online administration system in-house or out-house with a vendor to conduct the administration and score reporting.
- Standardize directions, administration procedures, and a narrow window for completion of your student rating scale and other measures of teaching effectiveness.

Correspondence: R.A. Berk, Johns Hopkins University, 10971 Swansfield Road, Columbia, MD 21044, USA. Tel: +1 410 9407118; fax: +1 206 3091618; email: rberk1@jhu.edu

ISSN 0142–159X print/ISSN 1466–187X online/13/010015–12 © 2013 Informa UK Ltd.
DOI: 10.3109/0142159X.2012.732247

faculty and administrators more than others. This contribution is an attempt to decrease some of that percolating and popping.

Trouble-shooting flashpoints

If you are currently using any instrument to measure teaching performance in face-to-face, blended/hybrid, or online courses, then you are probably struggling with one or more flashpoints. This article is a "consumer's guide to trouble-shooting these flashpoints." The motto of this article is: "Get to the flashpoint and the solution."

This is the inauguration of my new PBW series on *problem-based writing*. Your problems are the foci of my writing. The structure of each section will be governed by the PBW perspective:

(1) *Definition:* Each flashpoint will be succinctly defined.
(2) *Options:* The options available based on research and practice will be described.
(3) *Recommended Solution:* Specific, concrete recommendations for faculty and administrators will be proffered to move them to action.

There does not seem to be any short-cut, quick fix, or multi-level marketing scheme to improve the quality of teaching. Tackling these flashpoints head-on will hopefully be one positive step toward that improvement.

The *top five flashpoints* are: (1) student ratings vs. multiple sources of evidence; (2) sources of evidence vs. decisions: which come first?; (3) quality of "home-grown" rating scales vs. commercially-developed scales; (4) paper-and-pencil vs. online scale administration; and (5) standardized vs. unstandardized online scale administration. The first three relate to critical decisions about the sources of evidence chosen and the last two pertain to online scale administration issues.

Top five flashpoints

Student ratings vs. multiple sources of evidence

> **FLASHPOINT 1:** *Student rating scales have dominated as the primary or, usually, the only measure of teaching effectiveness in medical schools/colleges and universities worldwide and in a few remote planets. This state of practice is contrary to the advice of a cadre of experts and the limitations of student input to comprehensively evaluate teaching effectiveness. Several other measures should be used in conjunction with student ratings.*

Student ratings. Historically, student rating scales have been the primary measure of teaching effectiveness for the past 50 years. Students have had a critical role in the teaching–learning feedback system. The input from their ratings in summative decision making has been recommended on an international level (Strategy Group 2011; Surgenor 2011).

There are nearly 2000 references on the topic (Benton & Cashin 2012) with the first journal article published 90 years ago (Freyd 1923). There is more research and experience in higher education with student ratings than with all of the other measures of teaching effectiveness combined (Berk 2005, 2006). If you need to be brought up to speed quickly with the research on student ratings, check out these up-to-date reviews (Gravestock & Gregor-Greenleaf 2008; Benton & Cashin 2012; Kite 2012).

Unfortunately, in medical/healthcare education, student ratings have not received the same level of research attention. There is only a sprinkling of studies over the last 20 years (e.g., Hoeks & van Rossum 1988; Jones & Froom 1994; Mazor et al. 1999; Elzubeir & Rizk 2002; Barnett et al. 2003; Kidd & Latif 2004; Pierre et al. 2004; Turhan et al. 2005; Maker et al. 2006; Ahmady et al. 2009; Barnett & Matthews 2009; Berk 2009a; Chenot et al. 2009; Donnon et al. 2010; Boerboom et al. 2012; Stalmeijer et al. 2010). There is far more research on peer observation (e.g., Berk et al. 2004; Siddiqui et al. 2007; Wellein et al. 2009; DiVall et al. 2012; Pattison et al. 2012; Sullivan et al. 2012). There are also a few qualitative studies that are peripherally related (Stark 2003; Steinert 2004; Martens et al. 2009; Schiekirka et al. 2012).

With this volume of scholarly productivity and practice in academia worldwide, student ratings seem like the solution to assessing teaching effectiveness in medical/healthcare education and higher education in general. So, what is the problem?

Limitations of student ratings. As informative as student ratings can be about teaching, there are numerous *behaviors and skills defining teaching effectiveness which students are NOT qualified to rate*, such as a professor's knowledge and content expertise, teaching methods, use of technology, course materials, assessment instruments, and grading practices (Cohen & McKeachie 1980; Calderon et al. 1996; d'Apollonia & Abrami 1997a; Ali & Sell 1998; Green et al. 1998; Hoyt & Pallett 1999; Coren 2001; Ory & Ryan 2001; Theall & Franklin 2001; Marsh 2007; Svinicki & McKeachie 2011). Students can provide feedback at a certain level in each of those areas, but it will take peers and other qualified professionals to rate those skills in depth. *BOTTOM LINE: Student ratings from well-constructed scales are a necessary, but not sufficient, source of evidence to comprehensively assess teaching effectiveness.*

Student ratings provide only one portion of the information needed to infer teaching effectiveness. Yet, that is pretty much all that is available at most institutions. When those ratings alone are used for decision making, they will be incomplete and biased. Without additional evidence of teaching effectiveness, *student ratings can lead to incorrect and unfair career decisions about faculty that can affect their contract renewal, annual salary increase, and promotion and tenure.*

It is the process of *evaluation* or *assessment* that permits several sources of appropriate evidence to be collected for the purpose of decision making. *Assessment* is a "systematic method of obtaining information from [scales] and other sources, used to draw inferences about characteristics of people, objects, or programs," according to the US *Standards for Educational and Psychological Testing* (AERA, APA, & NCME Joint Committee on Standards 1999, p. 272). *Student ratings represent one measure and just one source of information in that process.*

Multiple sources of evidence. Over the past decade, there has been a trend toward augmenting student ratings with other data sources of teaching effectiveness. Such sources can serve to broaden and deepen the evidence base used to assess courses and the quality of teaching (Theall & Franklin 1990; Braskamp & Ory 1994; Hoyt & Pallett 1999; Knapper & Cranton 2001; Ory 2001; Cashin 2003; Berk 2005, 2006; Seldin 2006; Arreola 2007; Theall & Feldman 2007; Gravestock & Gregor-Greenleaf 2008; Benton & Cashin 2012). In fact, several comprehensive models of "faculty evaluation" have been proposed (Centra 1993; Braskamp & Ory 1994; Berk 2006, 2009a; Arreola 2007; Gravestock & Gregor-Greenleaf 2008), which include multiple sources of evidence with some models attaching greater weight to student and peer ratings and less weight to self-, administrator, and alumni ratings, and other sources. All of these models are used to arrive at formative and summative decisions.

15 Sources. There are 15 potential sources of evidence of teaching effectiveness: (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. Berk (2006) described several major characteristics of each source, including type of measure needed to gather the evidence, the person(s) responsible for providing the evidence (students, peers, external experts, mentors, instructors, or administrators), the person or committee who uses the evidence, and the decision(s) typically rendered based on that data (formative, summative, or program). He also critically examined the value and contribution of these sources for teaching effectiveness based on the current state of research and practice. His latest recommendations will be presented in Flashpoint 2.

Triangulation. Much has been written about the merits and shortcomings of these various sources of evidence (Berk 2005, 2006). Put simply: *There is no perfect source or combination of sources.* Each source can supply unique information, but also is fallible, usually in a way different from the other sources. For example, the unreliability and biases of peer ratings are not the same as those of student ratings; student ratings have other weaknesses. By drawing on three or more different sources of evidence, you can leverage the *strengths of each source to compensate for weaknesses of the other sources,* thereby converging on a decision about teaching effectiveness that is more accurate and reliable than one based on any single source (Appling et al. 2001). This notion of *triangulation* is derived from a compensatory model of decision making.

Given the complexity of measuring the act of teaching in a real-time classroom environment or online course, it is reasonable to expect that *multiple sources can provide a more accurate, reliable, and comprehensive picture of teaching effectiveness than just one source.* However, the decision maker should integrate the information from only those sources for which validity evidence is available (see Standard 14.13). The quality of the sources chosen should be beyond reproach, according to the *Standards* (AERA, APA, & NCME Joint Committee on Standards 1999).

Since there is not enough experience with multiple sources, there is a scarcity of empirical evidence to support the use of any particular combination of sources (e.g., Barnett et al. 2003; Stalmeijer et al. 2010; Stehle et al. 2012). There are a few surveys of the frequency of use of individual sources (Seldin 1999; Barnett & Matthews 2009). Research is needed on various combinations of measures for different decisions to determine "best practices."

Recommendations. All experts on faculty evaluation recommend multiple sources of evidence to assess teaching effectiveness. Beyond student ratings, is it worth the extra effort, time, and cost to develop the additional measures suggested in this section? Just what new information do you have to gain?

As those instruments are being built, it should become clear that they are intended to measure different teaching behaviors that contribute to teaching effectiveness. Each measure should bite off a separate chunk of behaviors. They should be designed to be complementary, not redundant, although there may be justification for some overlap for corroboration.

There is even research evidence on the relationships between student ratings and several other measures to support their complementarity. Benton and Cashin's (2012) research review reported the following validity coefficients with student ratings: trained *observers* (0.50 with global ratings), *self* (0.30–0.45), *alumni* (0.54–0.80), and *administrators* (0.47–0.62; 0.39 with global ratings). Since 0.50 is only 25% explained variance and 75% unexplained or new information, these coefficients suggest a lot of insight can be gained using observers', self, and administrators' ratings as sources of evidence.

Sources of evidence vs. decisions: Which come first?

> **FLASHPOINT 2:** *Rating scales are typically administered and then confusion occurs over what to do with the results and how to interpret them for specific decisions. A better strategy would be to do exactly the opposite of that practice. Spin your head around 180°, exorcist style. The decision should drive the selection of the appropriate sources of evidence, the types of data needed for the decision, and the design of the report form. Custom tailor the sources, data, and form to fit the decision. The information and format of the evidence a professor needs to improve his or her teaching are very different from that required by a department chair or associate dean for annual review (contract renewal or merit pay) or by a faculty committee for promotion and tenure review. The sources of evidence and formats of the reports can either hinder or facilitate the decision process.*

Types of decisions. According to Seldin (1999), *teaching is the major criterion (98%) in assessing overall faculty performance* in liberal arts colleges compared to student advising (64%), committee work (59%), research (41%), publications (31%), and public service (24%). Although these figures may not hold up in research universities and, specifically, in medical schools/colleges, teaching didactic, and/or clinical courses is still a critical job requirement and criterion on which most faculty members are assessed.

There are two types of *individual decisions* in faculty assessment with which you may already be familiar in the context of student assessment, plus one decision about *programs*:

(1) *Formative decisions*. These are decisions faculty make to *improve and shape the quality of their teaching*. It is based on evidence of teaching effectiveness they gather to plan and revise their teaching semester after semester. This evidence and the subsequent adjustments in teaching can occur anytime during the course, so the students can benefit from those changes, or after the course in preparation for the next course.

(2) *Summative decisions*. These decisions are rendered by the administrative-type person who controls a professor's destiny and future in higher education. This individual is usually the dean, associate dean, program director, or department head or chair. This administrator uses evidence of a professor's teaching effectiveness along with other evidence of research, publications, clinical practice, and service to *"sum up" his or her overall performance or status to decide about contract renewal or dismissal, annual merit pay, teaching awards, and promotion and tenure*.

Although promotion and tenure decisions are often made by a faculty committee, a letter of recommendation by the dean is typically required to reach the committee for review. These summative decisions are *high-stakes, final employment decisions* reached at different points in time to determine a professor's progression through the ranks and success as an academician.

(3) *Program decisions*. Several sources of evidence can also be used for program decisions, as defined in the *Program Evaluation Standards* by the US Joint Committee on Standards for Educational Evaluation (Yarbrough et al. 2011). They relate to the *curriculum, admissions and graduation requirements, and program effectiveness*. They are NOT individual decisions; instead, they *focus on processes and products*. The evidence usually is derived from various types of faculty and student input and employers' performance appraisal of students. It is also collected to provide documentation to satisfy the criteria for accreditation review.

Matching sources of evidence to decisions. The challenge is to pick the most appropriate and highest quality sources of evidence for the specific decision to be made; that is, match the sources to the decision. The decision drives your choices of evidence. Among the aforementioned 15 sources of evidence of teaching effectiveness, here are my best picks based on the literature for formative, summative, and program decisions:

Formative decisions

- student ratings,
- peer observations,
- peer review of course materials,
- external expert ratings,
- self-ratings,
- videos,
- student interviews, and
- mentor's advice.

Summative decisions (annual review for contract renewal and merit pay)

- student ratings,
- self-ratings,
- teaching scholarship,
- administrator ratings,
- teaching portfolio (for several courses over the year),
- peer observation (report written expressly for summative decision),
- peer review of course materials (report written expressly for summative decision), and
- mentor's review (progress report written expressly for summative decision).

Summative decisions (promotion and tenure)

- student ratings,
- self-ratings,
- teaching scholarship,
- administrator ratings,
- teaching portfolio (across several years' courses),
- peer review (written expressly for summative decision), and
- mentor's review (progress report written expressly for summative decision).

Program decisions

- Student ratings
- Exit and alumni ratings
- Employer ratings

The multiple sources identified for each decision can be configured into the *360° multisource feedback (MSF) model* of assessment (Berk 2009a, 2009b) or other model for accreditation documentation of teaching assessment. The sources for each decision may be added gradually to the model. This is an on-going process for your institution.

Recommendations. So now that you have seen my picks, which sources are you going to choose? So many sources, so little time! Which sources are already available in your department? What is the quality of the measures used to provide evidence of teaching effectiveness? Are the faculty stakeholders involved in the current process?

You have some decisions to make. Where do you begin? Here are a few suggestions:

(1) *Start with student ratings*. Consider the content and quality of your current scale and determine whether it needs a minor or major tune-up for the decisions being made.

(2) *Review the other sources of evidence* with your faculty to decide the next steps. Which sources will your faculty embrace which reflect best practices in

teaching? Weigh the pluses and minuses of the different sources.

(3) *Decide which combination of sources is best* for your faculty. Identify which sources should be used for both formative and summative decisions, such as self- and peer ratings, and which sources should be used for one type of decision but not the other, such as administrator ratings and teaching portfolio.

(4) *Map out a plan to build those sources*, one at a time, to create an assessment model for each decision (see Berk 2009a).

Whatever combination of sources you choose to use, take the time and make the effort to design the scales, administer the scales, and report the results appropriately. *The accuracy of faculty assessment decisions depends on the integrity of the process and the validity and reliability of the multiple sources of evidence you collect.* This endeavor may seem rather formidable, but, keep in mind, you are not alone in this process. Your colleagues at other institutions are probably struggling with the same issues. Maybe you could pool resources.

Quality of ''home-grown'' rating scales vs. commercially-developed scales

> **FLASHPOINT 3:** *Many of the rating scales developed by faculty committees in medical schools/colleges and universities do not meet even the most basic criteria for psychometric quality required by professional and legal standards. Most of the scales are flawed internally, administered incorrectly, and rarely is there any evidence of score reliability and validity. The serious concern is that decisions about the careers of faculty are being made with these instruments.*

Quality control. Researchers have reviewed the quality of student rating scales used by colleges and universities throughout the US and Canada (Berk 1979, 2006; Franklin & Theall 1990; d'Apollonia & Abrami 1997b, 1997c; Seldin 1999; Theall & Franklin 2000; Abrami 2001; Franklin 2001; Ory & Ryan 2001; Arreola 2007; Gravestock & Gregor-Greenleaf 2008). The instruments are either commercially developed scales with pre-designed reporting forms or "home-grown," locally constructed measures built usually by faculty committees. The former exhibit the quality control of the company that developed the scales and reports, such as Educational Testing Service and The IDEA Center (see Flashpoint 4); the latter have no consistency in the development process and rarely any formal procedures for controlling psychometric quality.

Quality of "home-grown" scales. That lack of quality control may very well extend to institutions worldwide. It could be *due to a lack of commitment, importance, accountability, or interest; inappropriate personnel without the essential skills; or limited resources.* No one knows for sure. Regardless of the reason, the picture is ugly.

Reviewers of practices at institutions in North America have found the following problems with "home-grown" scales:

- poor or no specifications of teaching behaviors,
- faulty items (statements and anchors),
- ambiguous or confusing directions,
- unstandardized administration procedures,
- inappropriate data collection, analysis, and reporting,
- no adjustments in ratings for extraneous factors,
- no psychometric studies of score reliability and validity, and
- no guidelines or training for faculty and administrators to use the results correctly for appropriate decisions.

Does the term *psychometrically putrid* summarize current practices? How does your scale stack up against those problems? Fertilizer-wise, "home-grown" scales are not growing. Their development is arrested. They are more like "Peter Pan scales."

The potential negative consequences of using faulty measures to make biased and unfair decisions to guide teaching improvement and faculty careers can be devastating. Moreover, this assessment only addresses the quality of student rating scales. What would be the quality of peer observations, self-ratings, and administrator ratings and their interpretations? Serious attention needs to be devoted to the quality control of all "home-grown" scales.

From a broader perspective, *poor quality scales violates US testing/scaling standards* according to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME Joint Committee on Standards 1999), *Personnel Evaluation Standards* (Joint Committee on Educational Evaluation Standards 2009), and the US Equal Employment Opportunity Commission's (EEOC) *Uniform Guidelines on Employee Selection Procedures* (US Equal Employment Opportunity Commission 2010). The psychometric requirements for instruments used for summative "employment" decisions about faculty are rigorous and appropriate for their purposes.

Recommendations. This issue reduces to the leadership and the composition of the faculty committee that accepts the responsibility to develop the scales and reports and/or the external consultant or vendor hired to guide the development process. *The psychometric standards for the construction, administration, analysis, and interpretation of scales must be articulated and guided by professionals trained in those standards* (AERA, APA, & NCME Joint Committee on Standards 1999). As Flashpoint 2 emphasized, if the committee does not contain one or more professors with expertise in psychometrics, then it should be ashamed of itself. That is a prescription for putridity and the previous problem list. Reviewers rarely found any one with these skills on the committees of the institutions surveyed.

It is also recommended that *all faculty members* be given workshops on item writing and scale structure. In the development process, they *will be reviewing, selecting, critiquing, adapting, and writing items.* Even if faculty are excellent test item writers, that does not mean they can write scale items.

The structure and criteria for writing scale items are very different from test items (Berk 2006), not difficult, just different. Even with commercially developed instruments, *professors are usually given the option to add up to 10 course-specific items*; in other words, *they will need to write items*. Rules for writing scale items are available in references on scale construction (Netemeyer et al. 2003; Dunn-Rankin et al. 2004; Streiner & Norman 2008; Berk 2006; deVellis 2012).

Paper-and-pencil vs. online scale administration

FLASHPOINT 4: *The battle between paper-and-pencil versus online administration of student rating scales is still being fought in medical schools and on many campuses worldwide. Despite an international trend and numerous advantages and improvements in online systems over the past decade, there are faculty who still dig their heels in and institutions that have resisted the conversion. Much has been learned about how to increase response rates, which is a flashpoint by itself, and how to overcome many of the deterrents to adopting an online system. Online administration, analysis, and reporting can be executed in-house or by an out-house vendor specializing in that processing.*

Comparison of paper-and-pencil and online administration. A detailed examination of the advantages and disadvantages of the two modes of administration according to 15 key factors has been presented by Berk (2006). There are major differences between them. Although it was concluded that both are far from perfect, the *benefits of the online mode and the improvements in the delivery system with the research and experiences over the past few years exceed the pluses of the paper-based mode*. Furthermore, most Net Geners do not know what a pencil is. Unless it is an iPencil, it is not on their radar or part of their mode.

The benefits of the online mode include *ease of administration, administration flextime, low cost, rapid turnaround time for results, ease of scale revision, and higher quality and greater quantity of unstructured responses* (Sorenson & Johnson 2003; Anderson et al. 2005; Berk 2006; Liu 2006; Heath et al. 2007). Students' concerns with lack of anonymity, confidentiality of ratings, inaccessibility, inconvenience, and technical problems have been eliminated at many institutions. Faculty resistance issues of low response rates and negative bias and lower ratings than paper-based version have been addressed (Berk 2006). Two major topics that still need attention are lack of standardization (Flashpoint 5) and response bias, which tends to be the same for both paper and online.

Three online delivery options. Online administration, scoring, analysis, and reporting of student ratings can be handled in three ways: (1) in-house by the department of computer services, IT, or equivalent unit; (2) out-house by a vendor that provides all delivery services for the institution's "home-grown" scale; or (3) out-house by a vendor that provides all services, plus their own scale or a scale you create from their catalog of items. These options are listed in order of increasing cost. Depending on in-house resources, it is possible to execute the entire processing in a very cost-effective manner. Alternatively, estimates from a variety of vendors should be obtained for the out-house options.

(1) *In-house administration.* If you have developed or plan to *develop your own scale*, you should consider this option. Convene the key players who can make this happen, including administrators and staff from IT or computer services, faculty development, and a testing center, plus at least one measurement expert. *A discussion of scale design, scoring, analysis, report design, and distribution can determine initially whether the resources are available to execute the system.* Once a preliminary assessment of the resources required has been completed, costs should be estimated for each phase. A couple of meetings can provide enough information to consider the possibility.

Your in-house system components, products, and personnel can then be compared to the two options described next. As you go shopping for an online system, at least *you will have done your homework and be able to identify what the commercial vendors offer*, including qualitative differences, that you cannot execute yourself. Although the cost could be the deal-breaker, you will know all the options available to make an informed final decision. Further, you can always change your system if your stocks plummet, the in-house operation has too many bumps that cannot be squished and ends up in Neverland, or the commercial services do not deliver as promised.

(2) *Vendor administration with "home-grown" scale.* If outsourcing to a vendor is your preference or you just want to explore this option, but you want to *maintain control over your own scale content and structure*, there are certain vendors that can online your scale. For some strange reason, they are all located in Madagascar. Kidding. They include CollegeNET (What Do You Think?), ConnectEDU (courseval), and IOTA Solutions (MyClassEvaluation). They will administer your scale online, perform all analyses, and generate reports for different decision makers. Thoroughly compare all of their components with yours. Evaluate the pluses and minuses of each package.

Make sure to investigate the compatibility of the packages with your course management system. The choice of the system is crucial to provide the anonymity for students to respond, which can boost response rates (Oliver & Sautter 2005). Most of the vendors' packages are compatible with Blackboard, WebCT, Moodle, Sakai, and other campus portal systems.

There are even *free online survey providers*, such as Zoomerang (MarketTools 2006), which can be *used easily by any instructor without a course management system* (Hong 2008). Other online survey software, both free and pay, has been reviewed by Wright (2005). There are specific advantages and disadvantages of the different packages, especially with regard to rating scale structure and reporting score results (Hong 2008). This is a viable online option worth investigating for formative feedback.

(3) *Vendor administration and rating scale.* If you want a vendor to supply the rating scale and all of the delivery services, there are several commercial student rating systems you should consider. Examples include *Online Course Evaluation, Student Instructional Report II, Course/Instructor Evaluation Questionnaire, IDEA Student Ratings of Instruction, Student Evaluation of Educational Quality, Instructional Assessment System, and Purdue Instructor Course Evaluation Service.* Sample forms and lists of services with prices are given on the websites for these scales.

This is the *simplest solution to the student rating scale online system*: Just go buy one. The seven packages are designed for you, Professor Consumer. The items are professionally developed; the scale has usually undergone extensive psychometric analyses to provide evidence of reliability and validity; and there are a variety of services provided, including the scale, online administration, scanning, scoring, and reporting of results in a variety of formats with national norms. For some, you can access a specimen set of rating scales and report forms online. All of the vendors provide a list of services and prices on their websites.

Carefully shop around to *find the best fit for your faculty and administrator needs and institutional culture*. The packages vary considerably in scale design, administration options, report forms, norms, and, of course, cost.

Comparability of paper-and-pencil and online ratings. Despite all of the differences between paper-based and online administrations and the contaminating biases that afflict the ratings they produce, researchers have found consistently that *online students and their in-class counterparts rate courses and instructors similarly* (Layne et al. 1999; Spooner et al. 1999; Waschull 2001; Carini et al. 2003; Hardy 2003; McGee & Lowell 2003; Dommeyer et al. 2004; Avery et al. 2006; Benton et al. 2010b; Venette et al. 2010; Perrett 2011; Stowell et al. 2012). The ratings on the *structured items* are not systematically higher or lower for online administrations. The correlations between online and paper-based global item ratings were 0.84 (overall instructor) and 0.86 (overall course) (Johnson 2003).

Although the ratings for online and paper are not identical, with more than 70% of the variance in common, any differences in ratings that have been found are small. Further, interrater reliabilities of ratings of individual items and item clusters for both modalities were comparable (McGee & Lowell 2003), and so were the underlying factor structures (Layne et al. 1999; Leung & Kember 2005). All of these similarities were also found in comparisons between face-to-face and online courses, although response rates were slightly lower in the online courses (Benton et al. 2010a).

Alpha total scale (18 items) reliabilities were similar for paper-based (0.90) and online (0.88) modes when all items appeared on the screen (Peer & Gamliel 2011). Slightly lower coefficients (0.74–0.83) for online displays of one, two, or four items only on the screen were attributable to response bias (Gamliel & Davidovitz 2005; Berk 2010; Peer & Gamliel 2011).

The one exception to the above similarities is the *unstructured items*, or open-ended comment section. The research has indicated that the flexible time permitted to the onliners usually, but not always, yields *longer, more frequent and thoughtful comments than those of in-class respondents* (Layne et al. 1999; Ravelli 2000; Johnson 2001, 2003; Hardy 2002, 2003; Anderson et al. 2005; Donovan et al. 2006; Venette et al. 2010; Morrison 2011). Typing the responses is reported by students to be easier and faster than writing them, plus it *preserves their anonymity* (Layne et al. 1999; Johnson 2003).

Recommendations. Weighing all of the pluses and minuses in this section strongly suggests that the *conversion from a paper-based to online administration system seems worthy of serious consideration by medical schools/colleges and every other institution of higher education using student ratings.* When the concerns of the online approach are addressed, its benefits for face-to-face, blended/hybrid, and online/distance courses far outweigh the traditional paper-based approach. (*NOTE:* Online administration should also be employed for alumni ratings and employer ratings. The costs for these ratings will be a small fraction of the cost of the student rating system.)

Standardized vs. unstandardized online scale administration

FLASHPOINT 5: *Standardized administration procedures for any measure of human or rodent behavior are absolutely essential to be able to interpret the ratings with the same meaning for all individuals who completed the measure. Student rating scales are typically administered online at the end of the semester without regard for any standardization or controls. There doesn't seem to be any sound psychometric reasons for why the administrations are scheduled the way they are. This is, perhaps, the most neglected issue in the literature and in practice.*

Importance of standardization. A significant amount of attention has been devoted to establishing standardized times, conditions, locations, and procedures for administering in-class tests and clinical measures, such as the OSCE, as well as out-of-class admissions, licensing, and certification tests. National standards for testing practices require this standardization to assure that students take tests under identical conditions so their scores can be interpreted in the same way, they are comparable from one student or group to another, and they can be compared to norms (AERA, APA, & NCME Joint Committee on Standards 1999).

Unfortunately, *standardization has been completely neglected in the faculty evaluation literature for the administration of online student rating scales* (Berk 2006). This topic was only briefly mentioned in a recent review of the student ratings research (Addison & Stowell 2012). Although the inferences drawn from the scale scores and other measures of teaching effectiveness require the same administration precision as tests, procedures to assure scores will have the same meaning from students completing the scales at the end of the semester have not been addressed in research and practice. Typically, students are given notice that they have 1 or 2 weeks to complete the student ratings form with the deadline before or after the final exam/project.

Confounding uncontrolled factors. Since students can complete online rating scales during their discretionary time, there is *no control over the time, place, conditions, or any situational factors* under which the self-administrations occur (Stowell et al. 2012). Most of these factors were controlled with the paper-and-pencil, in-class administration by the instructor or a student appointed to handle the administration.

In fact, in the online mode, there is *no way to insure that the real student filled out the form or did not discuss it with someone who already did.* It could be a roommate, partner, avatar, alien, student who has never been to class doing a favor in exchange for a pizza, alcohol, or drugs, or all of the preceding. Any of those substitutes would result in *fraudulent ratings* (Standard 5.6). Bad, bad ratings! Although there is no standardization of the actual administration, at least the written directions given to all students can be the same. Therefore, the procedures that the students follow should be similar if they read the directions.

Timing of administration. The timing of the administration can also markedly affect the ratings. For example, if some students complete the scale before the final review and final exam, on the day of the final, or after the exam, their feelings about the instructor/course can be very different. *Exposure to the final exam alone can significantly affect ratings*, particularly if there are specific items on the scale measuring testing and evaluation methods. It could be argued that the final should be completed in order to provide a true rating of all evaluation methods.

Despite a couple of "no difference" studies of paper-and-pencil administrations almost 40 years ago (Carrier et al. 1974; Frey 1976) and one study examining final exam day administration (Ory 2001), which produced lower ratings, *there does not seem to be any agreement among the experts on the best time to administer online scales or on any specific standardization procedures* other than directions.

What is clear is that whatever time is decided must be the same for all students in all courses; otherwise, the ratings of these different groups of students will not have the same meaning. For example, faculty within a department should agree that all online administrations must be completed before the final or after, but not both. Faculty must decide on the best time to get the most accurate ratings. That decision will also affect the legitimacy of any comparison of the ratings to different norm groups.

Standards for standardization. So what is the problem with the lack of standardization? The ratings by students are assumed to be collected under identical conditions according to the same rules and directions. Standardization of the administration and environment provide a snapshot of how students feel at one point in time. Although their individual ratings will vary, they will have the same meaning. Rigorous procedures for standardization are required by the US *Standards for Educational and Psychological Testing* (AERA, APA, & NCME Joint Committee on Standards 1999).

Groups of students must be given identical instructions, which is possible, administered the scale under identical conditions, which is nearly impossible, to assure the comparability of their ratings (Standards 3.15, 3.19, and 3.20). Only then would the interpretation of the ratings and, in this case, the inferences about teaching effectiveness from the ratings be valid and reliable (Standard 3.19). In other words, *without standardization*, such as when every student fills out the scale willy-nilly at different times of the day and semester, in different places, under different conditions, using different procedures, *the ratings from student to student and professor to professor will not be comparable.*

Recommendations. Given the limitations of online administration, what can be done to *approximate some semblance of standardized conditions* or, at least, *minimize the extent to which the bad conditions contaminate the ratings*? Here are a few options extended from Berk's (2006) previous suggestions, listed from highest level of standardization and control to lowest level:

(1) *In-class administration before final for maximum control:* Set a certain time slot in class, just like the paper-and-pencil version, for students to complete the forms on their own PC/Mac, iPad, iPhone, iPencil, or other device. The professor should leave the room and have a student execute and monitor the process. Adequate time should be given for students to type comments for the unstructured section of the scale. (*NOTE:* Not recommended if there are several items or a subscale that measures course evaluation methods, since the final is part of those methods.)

(2) *Computer lab time slots before or after final:* Set certain time slots in the computer lab or an equivalent location during which students can complete the forms. The controls exercised in the previous option should be followed in the lab. If available, techie-type students should proctor the slots to eliminate distractions and provide technical support for any problems that arise.

(3) *One or two days before or after final at students' discretion:* This is the *most loosy-goosy option with the least control*, albeit, the most popular. Specify a *narrow window* within which the ratings must be completed, such as *one or two days after the final class and before the final exam*, or *one or two days after the exam before grades are submitted and posted.* This gives new meaning to "storm window."

Any of these three options will improve the standardization of current online administration practices beyond the typical 1- or 2-week bay window. Experience and research on these procedures will hopefully identify the confounding variables that can affect the online ratings. Ultimately, concrete guidelines to assist faculty in deciding on the most appropriate administration protocol will result.

Top five recommendations

After ruminating over these flashpoints, it can be concluded that there are a variety of options within the reach of every medical school/college and institution of higher education to improve its current practices with its source(s) of evidence and administration procedures. Everyone is wrestling with these issues and, although more research is needed to test the options, there are tentative solutions to these problems. As experience and research continue to accumulate, even better solutions will result.

There is a lot of activity and discourse on these flashpoints because we know that all of the summative decisions about faculty will be made with or without the best information available. Further, professors who are passionate about teaching will also seek out sources of evidence to guide their improvement.

The contribution of this PBW article rests on the value and usefulness of the recommendations that you can convert into action. Without action, the recommendations are just dead words on a page. *Your TAKE-AWAYS are the concrete action steps you choose to implement to improve the current state of your teaching assessment system.*

Here are the top five recommendations framed in terms of action steps:

(1) polish your student rating scale, but also start building additional sources of evidence, such as self, peer, and mentor scales, to assess teaching effectiveness;

(2) match your highest quality sources to the specific formative and summative decisions using the 360° MSF model;

(3) review current measures of teaching effectiveness with your faculty and plan specifically how you can improve their psychometric quality;

(4) design an online administration system in-house or out-house with a vendor to conduct the administration and score reporting for your own student rating scale or the one it provides; and

(5) standardize directions, administration procedures, and a narrow window for completion of your student rating scale and other measures of teaching effectiveness.

Taking action on these five can yield major strides in improving the practice of assessing teaching effectiveness and the fairness and equity of the formative and summative decisions made with the results. *Just how important is teaching in your institution?* Your answer will be expressed in your actions. What can you contribute to make it better than it is ever been? That is my challenge to you.

Notes on contributor

RONALD A. BERK, PhD, is Professor Emeritus, Biostatistics and Measurement, and former Assistant Dean for Teaching at the Johns Hopkins University, where he taught for 30 years. He has presented 400 keynotes/workshops and published 14 books, 165 journal articles, and 300 blogs. His professional motto is: “Go for the Bronze!”

Declaration of interest: The author report no conflicts of interest. The author alone is responsible for the content and writing of the article.

References

Medical/healthcare education

Ahmady S, Changiz T, Brommels M, Gaffney FA, Thor J, Masiello I. 2009. Contextual adaptation of the Personnel Evaluation Standards for assessing faculty evaluation systems in developing countries: The case of Iran. BMC Med Educ 9(18), DOI: 10.1186/1472-6920-9-18.

Anderson HM, Cain J, Bird E. 2005. Online student course evaluations: Review of literature and a pilot study. Am J Pharm Educ 69(1):34–43. Available from http://web.njit.edu/~bieber/pub/Shen-AMCIS2004.pdf.

Appling SE, Naumann PL, Berk RA. 2001. Using a faculty evaluation triad to achieve evidenced-based teaching. Nurs Health Care Perspect 22:247–251.

Barnett CW, Matthews HW. 2009. Teaching evaluation practices in colleges and schools of pharmacy. Am J Pharm Educ 73(6).

Barnett CW, Matthews HW, Jackson RA. 2003. A comparison between student ratings and faculty self-ratings of instructional effectiveness. J Pharm Educ 67(4).

Berk RA. 2009a. Using the 360° multisource feedback model to evaluate teaching and professionalism. Med Teach 31(12):1073–1080.

Berk RA, Naumann PL, Appling SE. 2004. Beyond student ratings: Peer observation of classroom and clinical teaching. Int J Nurs Educ Scholarsh 1(1):1–26.

Boerboom TBB, Mainhard T, Dolmans DHJM, Scherpbier AJJA, van Beukelen P, Jaarsma ADC. 2012. Evaluating clinical teachers with the Maastricht clinical teaching questionnaire: How much ‘teacher’ is in student ratings? Med Teach 34(4):320–326.

Chenot J-F, Kochen MM, Himmel W. 2009. Student evaluation of a primary care clerkship: Quality assurance and identification of potential for improvement. BMC Med Educ 9(17), DOI: 10.1186/1472-6920-9-17.

DiVall M, Barr J, Gonyeau M, Matthews SJ, van Amburgh J, Qualters D, Trujillo J. 2012. Follow-up assessment of a faculty peer observation and evaluation program. Am J Pharm Educ 76(4).

Donnon T, Delver H, Beran T. 2010. Student and teaching characteristics related to ratings of instruction in medical sciences graduate programs. Med Teach 32(4):327–332.

Elzubeir M, Rizk D. 2002. Evaluating the quality of teaching in medical education: Are we using the evidence for both formative and summative purposes? Med Teach 24:313–319.

Hoeks TW, van Rossum HJ. 1988. The impact of student ratings on a new course: The general clerkship (ALCO). Med Educ 22(4):308–313.

Jones RF, Froom JD. 1994. Faculty and administration views of problems in faculty evaluation. Acad Med 69(6):476–483.

Kidd RS, Latif DA. 2004. Student evaluations: Are they valid measures of course effectiveness? J Pharm Educ 68(3).

Maker VK, Lewis MJ, Donnelly MB. 2006. Ongoing faculty evaluations: Developmental gain or just more pain? Curr Surg 63(1):80–84.

Martens MJ, Duvivier RJ, van Dalen J, Verwijnen GM, Scherpbier AJ, van der Vleuten. 2009. Student views on the effective teaching of physical examination skills: A qualitative study. Med Educ 43(2):184–191.

Mazor K, Clauser B, Cohen A, Alper E, Pugnaire M. 1999. The dependability of students’ rating of preceptors. Acad Med 74:19–21.

Pattison AT, Sherwood M, Lumsden CJ, Gale A, Markides M. 2012. Foundation observation of teaching project – A developmental model of peer observation of teaching. Med Teach 34(2):e36–e142.

Pierre RB, Wierenga A, Barton M, Branday JM, Christie CD. 2004. Student evaluation of an OSCE in paediatrics at the University of the West Indies, Jamaica. BMC Med Educ 4(22), DOI: 10.1186/1472-6920-4-22.

Schiekirka S, Reinhardt D, Heim S, Fabry G, Pukrop T, Anders S, Raupach T. 2012. Student perceptions of evaluation in undergraduate medical education: A qualitative study from one medical school. BMC Med Educ 12(45), DOI: 10.1186/1472-6920-12-45.

Siddiqui ZS, Jonas-Dwyer D, Carr SE. 2007. Twelve tips for peer observation of teaching. Med Teach 29(4):297–300.

Stalmeijer RE, Dolmans DH, Wolfhagen IH, Peters WG, van Coppenolle L, Scherpbier AJ. 2010. Combined student ratings and self-assessment provide useful feedback for clinical teachers. Adv Health Sci Educ Theory Pract 15(3):315–328.

Stark P. 2003. Teaching and learning in the clinical setting: A qualitative study of the perceptions of students and teachers. Med Educ 37(11):975–982.

Steinert Y. 2004. Student perceptions of effective small group teaching. Med Educ 38(3):286–293.

Sullivan PB, Buckle A, Nicky G, Atkinson SH. 2012. Peer observation of teaching as a faculty development tool. BMC Med Educ 12(26), DOI: 10.1186/1472-6920-12-26.

Turhan K, Yaris F, Nural E. 2005. Does instructor evaluation by students using a web-based questionnaire impact instructor performance? Adv Health Sci Educ Theory Pract 10(1):5–13.

Wellein MG, Ragucci KR, Lapointe M. 2009. A peer review process for classroom teaching. Am J Pharm Educ 73(5).

General higher education

Abrami PC. 2001. Improving judgments about teaching effectiveness using rating forms. In: Theall M, Abrami PC, Mets LA, editors. The student ratings debate: Are they valid? How can we best use them? (New Directions for Institutional Research, No. 109). San Francisco, CA: Jossey-Bass. pp 59–87.

Addison WE, Stowell JR. 2012. Conducting research on student evaluations of teaching. In: Kite ME, editor. Effective evaluation of teaching: A guide for faculty and administrators. pp 1–12. E-book [Accessed 6 June 2012] Available from the Society for the Teaching of Psychology website http://teachpsych.org/ebooks/evals2012/index.php.

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.

Ali DL, Sell Y. 1998. Issues regarding the reliability, validity and utility of student ratings of instruction: A survey of research findings. Calgary, AB: University of Calgary APC Implementation Task Force on Student Ratings of Instruction.

Arreola RA. 2007. Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system. 3rd ed. Bolton, MA: Anker.

Avery RJ, Bryan WK, Mathios A, Kang H, Bell D. 2006. Electronic course evaluations: Does an online delivery system influence student evaluations? J Econ Educ 37(1):21–37.

Benton SL, Cashin WE. 2012. Student ratings of teaching: A summary of research and literature (IDEA Paper no. 50). Manhattan, KS: The IDEA Center. [Accessed 8 April 2012] Available from http://www.theideacenter.org/sites/default/files/idea-paper_50.pdf.

Benton SL, Webster R, Gross A, Pallett W. 2010a. An analysis of IDEA Student Ratings of Instruction in traditional versus online courses (IDEA Technical Report no. 15). Manhattan, KS: The IDEA Center.

Benton SL, Webster R, Gross A, Pallett W. 2010b. An analysis of IDEA Student Ratings of Instruction using paper versus online survey methods (IDEA Technical Report no. 16). Manhattan, KS: The IDEA Center.

Berk RA. 1979. The construction of rating instruments for faculty evaluation: A review of methodological issues. J Higher Educ 50:650–669.

Berk RA. 2005. Survey of 12 strategies to measure teaching effectiveness. Int J Teac Learn Higher Educ 17(1):48–62. Available from http://www.isetl.org/ijtlthe.

Berk RA. 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.

Berk RA. 2009b. Beyond student ratings: "A whole new world, a new fantastic point of view." Essays Teach Excellence 20(1). Available from http://podnetwork.org/publications/teachingexcellence.htm.

Berk RA. 2010. The secret to the "best" ratings from any evaluation scale. J Faculty Dev 24(1):37–39.

Braskamp LA, Ory JC. 1994. Assessing faculty work: Enhancing individual and institutional performance. San Francisco, CA: Jossey-Bass.

Calderon TG, Gabbin AL, Green BP. 1996. Report of the committee on promoting evaluating effective teaching. Harrisonburg, VA: James Madison University.

Carini RM, Hayek JC, Kuh GD, Ouimet JA. 2003. College student responses to web and paper surveys: Does mode matter? Res Higher Educ 44(1):1–19.

Carrier NA, Howard GS, Miller WG. 1974. Course evaluations: When? J Educ Psychol 66:609–613.

Cashin WE. 2003. Evaluating college and university teaching: Reflections of a practitioner. In: Smart JC, editor. Higher education: Handbook of theory and research. Dordrecht, the Netherlands: Kluwer Academic Publishers. pp 531–593.

Centra JA. 1993. Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness. San Francisco: Jossey-Bass.

Cohen PA, McKeachie WJ. 1980. The role of colleagues in the evaluation of teaching. Improving College Univ Teach 28(4):147–154.

Coren S. 2001. Are course evaluations a threat to academic freedom? In: Kahn SE, Pavlich D, editors. Academic freedom and the inclusive university. Vancouver, BC: University of British Columbia Press. pp 104–117.

d'Apollonia S, Abrami PC. 1997a. Navigating student ratings of instruction. Am Psychol 52:1198–1208.

d'Apollonia S, Abrami PC. 1997b. Scaling the ivory tower, part 1: Collecting evidence of instructor effectiveness. Psychol Teach Rev 6:46–59.

d'Apollonia S, Abrami PC. 1997c. Scaling the ivory tower, part 2: Student ratings of instruction in North America. Psychol Teach Rev 6:60–76.

deVellis RF. 2012. Scale development: Theory and applications. 3rd ed. Thousand Oaks, CA: Sage.

Dommeyer CJ, Baum P, Hanna RW, Chapman KS. 2004. Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. Assess Eval Higher Educ 29(5):611–623.

Donovan J, Mader CE, Shinsky J. 2006. Constructive student feedback: Online vs. traditional course evaluations. J Interact Online Learn 5:283–296.

Dunn-Rankin P, Knezek GA, Wallace S, Zhang S. 2004. Scaling methods. Mahwah, NJ: Erlbaum.

Franklin J. 2001. Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. In: Lewis KG, editor. Techniques and strategies for interpreting student evaluations (Special issue) (New Directions for Teaching and Learning, No. 87). San Francisco, CA: Jossey-Bass. pp 85–100.

Franklin J, Theall M. 1990. Communicating student ratings to decision makers: Design for good practice. In: Theall M, Franklin J, editors. Student ratings of instruction: Issues for improving practice (Special issue) (New Directions for Teaching and Learning, No. 43). San Francisco, CA: Jossey-Bass. pp 75–93.

Frey PW. 1976. Validity of student instructional ratings as a function of their timing. J Higher Educ 47:327–336.

Freyd M. 1923. A graphic rating scale for teachers. J Educ Res 8(5):433–439.

Gamliel E, Davidovitz L. 2005. Online versus traditional teaching evaluation: Mode can matter. Assess Eval Higher Educ 30(6): 581–592.

Gravestock P, Gregor-Greenleaf E. 2008. Student course evaluations: Research, models and trends. Toronto, ON: Higher Education Quality Council of Ontario. E-book [Accessed 6 May 2012] Available from http://www.heqco.ca/en-CA/Research/Research%20Publications/Pages/Home.aspx.

Green BP, Calderon TG, Reider BP. 1998. A content analysis of teaching evaluation instruments used in accounting departments. Issues Account Educ 13(1):15–30.

Hardy N. 2002. Perceptions of online evaluations: Fact and fiction. Paper presented at the annual meeting of the American Educational Research Association, April 1–5 2002, New Orleans, LA.

Hardy N. 2003. Online ratings: Fact and fiction. In: Sorenson DL, Johnson TD, editors. Online student ratings of instruction (New Directions for Teaching and Learning, No. 96). San Francisco, CA: Jossey-Bass. pp 31–38.

Heath N, Lawyer S, Rasmussen E. 2007. Web-based versus paper and pencil course evaluations. Teach Psychol 34(4):259–261.

Hong PC. 2008. Evaluating teaching and learning from students' perspectives in their classroom through easy-to-use online surveys. Int J Cyber Soc Educ 1(1):33–48.

Hoyt DP, Pallett WH. 1999. Appraising teaching effectiveness: Beyond student ratings (IDEA Paper no. 36). Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development.

Johnson TD. 2001. Online student ratings: Research and possibilities. Invited plenary presented at the Online Assessment Conference, September, Champaign, IL.

Johnson TD. 2003. Online student ratings: Will students respond?. In: Sorenson DL, Johnson TD, editors. Online student ratings of instruction (New Directions for Teaching and Learning, no. 96). San Francisco, CA: Jossey-Bass. pp 49–60.

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.

Kite ME, editor. 2012. Effective evaluation of teaching: A guide for faculty and administrators. E-book [Accessed 6 June 2012] Available from the Society for the Teaching of Psychology website http://teachpsych.org/ebooks/evals2012/index.php.

Knapper C, Cranton P, editors. 2001. Fresh approaches to the evaluation of teaching (New Directions for Teaching and Learning, no. 88). San Francisco, CA: Jossey-Bass. pp 19–29.

Layne BH, DeCristoforo JR, McGinty D. 1999. Electronic versus traditional student ratings of instruction. Res Higher Educ 40(2):221–232.

Leung DYP, Kember D. 2005. Comparability of data gathered from evaluation questionnaires on paper through the Internet. Res Higher Educ 46:571–591.

Liu Y. 2006. A comparison of online versus traditional student evaluation of instruction. Int J Instr Technol Distance Learn 3(3):15–30.

MarketTools. 2006. Zoomerang: Easiest way to ask, fastest way to know. [Accessed 17 July 2012] Available from http://info.zoomerang.com.

Marsh HW. 2007. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In: Perry RP, Smart JC, editors. The scholarship of teaching and learning in higher education: An evidence-based perspective. Dordrecht, the Netherlands: Springer. pp 319–383.

McGee DE, Lowell N. 2003. Psychometric properties of student ratings of instruction in online and on-campus courses. In: Sorenson DL, Johnson TD, editors. Online student ratings of instruction (New Directions for Teaching and Learning, no. 96). San Francisco, CA: Jossey-Bass. pp 39–48.

Morrison R. 2011. A comparison of online versus traditional student end-of-course critiques in resident courses. Assess Eval Higher Educ 36(6):627–641.

Netemeyer RG, Bearden WO, Sharma S. 2003. Scaling procedures. Thousand Oaks, CA: Sage.

Oliver RL, Sautter EP. 2005. Using course management systems to enhance the value of student evaluations of teaching. J Educ Bus 80(4):231–234.

Ory JC. 2001. Faculty thoughts and concerns about student ratings. In: Lewis KG, editor. Techniques and strategies for interpreting student evaluations (Special issue) (New Directions for Teaching and Learning, No. 87). San Francisco, CA: Jossey-Bass. pp 3–15.

Ory JC, Ryan K. 2001. How do student ratings measure up to a new validity framework?. In: Theall M, Abrami PC, Mets LA, editors. The student ratings debate: Are they valid? How can we best use them? (Special issue) (New Directions for Institutional Research, 109). San Francisco, CA: Jossey-Bass. pp 27–44.

Peer E, Gamliel E. 2011. Too reliable to be true? Response bias as a potential source of inflation in paper and pencil questionnaire reliability. Practical Assess Res Eval 16(9):1–8. Available from http://pareonline.net/getvn.asp?v=16%n=9.

Perrett JJ. 2011. Exploring graduate and undergraduate course evaluations administered on paper and online: A case study. Assess Eval Higher Educ 1–9, DOI: 10.1080/02602938.2011.604123.

Ravelli B. 2000. Anonymous online teaching assessments: Preliminary findings. [Accessed 12 June 2012] Available from http://www.edrs.com/DocLibrary/0201/ED445069.pdf.

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

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

Seldin P, Associates, editors. 2006. Evaluating faculty performance: A practical guide to assessing teaching, research, and service. Bolton, MA: Anker. pp 201–216.

Sorenson DL, Johnson TD, editors. 2003. Online student ratings of instruction (New Directions for Teaching and Learning, no. 96). San Francisco, CA: Jossey-Bass.

Spooner F, Jordan L, Algozzine B, Spooner M. 1999. Student ratings of instruction in distance learning and on-campus classes. J Educ Res 92:132–140.

Stehle S, Spinath B, Kadmon M. 2012. Measuring teaching effectiveness: Correspondence between students' evaluations of teaching and different measures of student learning. Res Higher Educ. DOI: 10.1007/s11162-012-9260-9.

Stowell JR, Addison WE, Smith JL. 2012. Comparison of online and classroom-based student evaluations of instruction. Assess Eval Higher Educ 37(4):465–473.

Strategy Group. 2011. National strategy for higher education to 2030 (Report of the Strategy Group). Dublin, Ireland: Department of Education and Skills, Government Publications Office. [Accessed 17 July 2012] Available from http://www.hea.ie/files/files/DES_Higher_Ed_Main_Report.pdf.

Streiner DL, Norman GR. 2008. Health measurement scales: A practical guide to their development and use. 4th ed. New York: Oxford University Press.

Surgenor PWG. 2011. Obstacles and opportunities: Addressing the growing pains of summative student evaluation of teaching. Assess Eval Higher Educ 1–14, iFirst Article. DOI: 10.1080/02602938.2011.635247.

Svinicki M, McKeachie WJ. 2011. McKeachie's teaching tips: Strategies, research, and theory for college and university teachers. 13th ed. Belmont, CA: Wadsworth.

Theall M, Feldman KA. 2007. Commentary and update on Feldman's (1997) "Identifying exemplary teachers and teaching: Evidence from student ratings". In: Perry RP, Smart JC, editors. The teaching and learning in higher education: An evidence-based perspective. Dordrecht, the Netherlands: Springer. pp 130–143.

Theall M, Franklin JL. 1990. Student ratings in the context of complex evaluation systems. In: Theall M, Franklin JL, editors. Student ratings of instruction: Issues for improving practice (New Directions for Teaching and Learning, no. 43). San Francisco, CA: Jossey-Bass. pp 17–34.

Theall M, Franklin JL. 2000. Creating responsive student ratings systems to improve evaluation practice. In: Ryan KE, editor. Evaluating teaching in higher education: A vision for the future (Special issue) (New Directions for Teaching and Learning, no. 83). San Francisco, CA: Jossey-Bass. pp 95–107.

Theall M, Franklin JL. 2001. Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction?. In: Theall M, Abrami PC, Mets LA, editors. The student ratings debate: Are they valid? How can we best use them? (New Directions for Institutional Research, no. 109). San Francisco, CA: Jossey-Bass. pp 45–56.

US Equal Employment Opportunity Commission (EEOC). 2010. Employment tests and selection procedures. [Accessed 20 August 2012] Available from http://www.eeoc.gov/policy/docs/factemployment_procedures.html.

Venette S, Sellnow D, McIntire K. 2010. Charting new territory: Assessing the online frontier of student ratings of instruction. Assess Eval Higher Educ 35:101–115.

Waschull SB. 2001. The online delivery of psychology courses: Attrition, performance, and evaluation. Comput Teach 28:143–147.

Wright KB. 2005. Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. J Comput Mediated Commun 10(3). Available from http://jcmc.indiana.edu/vol10/issue3/wright.html.

Yarbrough DB, Shulha LM, Hopson RK, Caruthers FA. 2011. The program evaluation standards: A guide for evaluators and evaluation users. 3rd ed. Thousand Oaks, CA: Sage.