Start Spreading the News: Use Multiple Sources of Evidence to Evaluate Teaching*

By Ronald A. Berk

LAW & ORDER: ***S**tudent **R**atings of **I**nstruction*
In the higher education system, the learning environment is supported by two separate, yet mega-important, groups: the professors who find value in SRIs and those who do not. These are their stories.
Ka Chung!

Introduction

Student ratings have "grit" (Duckworth, 2016), as much as any inanimate object could possibly have. They have sustained blunt-force trauma over and over again, but remain on top of the heap as the dominant approach to evaluate teaching. They provide outrage-ready headlines to academicians. The topic serves as clickbait for *Inside Higher Ed*, *The Chronicle of Higher Education*, higher education journals, listservs, and blogs (http://studentevaluationsareworthless.blogspot.com/), and fodder for critics and doom and gloom naysayers shouting to impeach student ratings. Linse (2017) identified 50 of these articles over the past decade, most of which were negative and unsupported by research evidence. In recent years, aspersions of "worthlessness" have become commonplace (Barre, 2015a, 2015b; Berrett, 2015; Boring, Ottoboni, & Stark, 2016; Braga, Paccagnella, & Pellizzari, 2014; Burt, 2015; Gooblar, 2017; Kamenetz, 2014; Nilson, 2012; Quintana, 2017; Schuman, 2014; Sproule & Valsan, 2009; Stroebe, 2016; Uttl, White, & Gonzalez, 2017).

(*SIDEBAR:* Despite my reputation for writing on scholarly topics with scrupulous impartiality and unfettered fairness, student ratings can cause unwanted fettering. I will try to control my fettering. We now resume this introduction with the lyrics to "New York, New York.")

Start spreading the news: Student ratings are not leaving today or anytime soon. However, as the title of this article indicates, student ratings are not the only option to provide evidence in the evaluation of teaching. There is a broad range of alternatives to consider beyond student ratings in the delicate decision-making processes to improve teaching and determine the promotion and tenure of faculty. Yet, despite the constant barrage of attacks on the integrity, reliability, and validity of student ratings, their use in higher education is at an all-time high.

So what do student ratings actually contribute to decisions about teaching and faculty? Should they be abandoned? Should you focus on the other options? This article examines student ratings and 14 alternatives to guide your plans to evaluate teaching in your department. Let's start with a brief review of …

Student Ratings

The student rating scale has been the primary measure of teaching effectiveness for more than three-quarters of a century. Currently, student ratings are "always used" as a major source of information to evaluate teaching performance by 94.2% of four-year liberal arts colleges in the U.S. (Miller & Seldin, 2014). In fact, they have been the universal performance barometer in colleges and universities worldwide. Perceptible barometric changes occur in

*This article is dedicated to the memory of my dear friend, Mike Theall, PhD, whose contributions to the teaching evaluation literature significantly advanced research and improved practices with student ratings. As a past president of the Professional and Organization Development Network in Higher Education, he is greatly missed for his leadership, intellect, integrity, compassion, generosity, and fantastic sense of humor. I hope the contents and humor herein will honor him.

this meteorological metaphor as students exercise their critical role in the teaching-learning feedback system. Their input in formative and summative decision making has been recommended on an international level (Griffin & Cook, 2009; Strategy Group, 2011; Surgenor, 2011).

Research Evidence

More has been written on this topic in higher education than any other. To date, there are *nearly 3,000 references* to student ratings (Benton & Cashin, 2014), with the *first journal article published 95 years ago* (Freyd, 1923). There is more research on and experience with student ratings than all of the other measures of teaching effectiveness combined (Berk, 2006, 2013d, 2018). That is tankers of research. If you need to be brought up to speed quickly with the research on student ratings, check out these readily available reviews (Benton & Cashin, 2012, 2014; Gravestock & Gregor-Greenleaf, 2008; Hativa, 2014a, 2014b; Kite, 2012). (*NOTE:* For a parody of the history of student ratings, see Berk, 2013d.)

With student ratings at the top of the leaderboard accompanied by an impressive volume of scholarly products and practices in academia, you would think that they would be the ideal tool to evaluate teaching. So …

What's the Problem?

There are four major limitations to using ONLY student ratings for decision making: (1) students' limited qualifications as raters, (2) technical inadequacy and bias, (3) misuse of scales and misinterpretation of ratings, and (4) inadequate source of evidence for decision making. Let's examine the significance of these limitations.

Students' limited qualifications as raters. As informative as student ratings can be, there are numerous *behaviors and skills that define teaching that students are not qualified to rate*, such as a professor's knowledge and content expertise, learning outcomes, teaching methods, course design and organization, use of technology, quality of course materials, assessment instruments, and grading practices (Ali & Sell, 1998; Benton & Li, 2017; Calderon, Gabbin, & Green, 1996; Cashin, 1989; Cohen & McKeachie,1980; Coren, 2001; d'Apollonia & Abrami, 1997; Green, Calderon, & Reider, 1998; Hoyt & Pallett, 1999; Keig & Waggoner, 1994; Marsh, 2007; Ory & Ryan, 2001; Svinicki & McKeachie, 2014; Theall, n.d.).

What's left that students *can* legitimately rate? They can provide feedback at a certain level in most of those areas, but it will take peers and other qualified professionals to rate those skills in depth. There are so many teaching behaviors to measure. Students should answer only those items that are directly within their purview of expertise and behaviors they have observed or experienced throughout the course.

"Why is that so important?" Good question. In employment decisions, a certain validity standard (or criterion) must be met for the scale being used: Each measure should be completed by those individuals (students, instructor, other faculty, administrators, or employers) who are in the best position to provide the most accurate information; otherwise, that information, in this case, rating scale scores, may be invalid or biased.

Technical inadequacy of ratings and bias. Student rating scales are constantly being criticized on technical grounds. Most home-grown forms, especially, do not meet basic psychometric specifications for employment decisions (AERA, APA, & NCME Joint Committee on Standards, 2014; U.S. EEOC, 2010) compared to those developed commercially (Berrett, 2015; Uttl et al., 2017; Wieman, 2015). The validity of the ratings has been challenged by Nilson (2012) and Uttl et al. (2017) in three areas: (1) weak relationship between student ratings and learning, (2) sources of bias in the ratings, including professor's charisma, physical attractiveness, personality, gender, age, race/ethnicity, rank, and class length (also see Addison & Stowell, 2012; Basow & Martin, 2012; Benton & Ryalls, 2016; Boring, 2017; Boring et al., 2016; Li & Benton, 2017; Linse, 2017; Macnell, Driscoll, & Hunt, 2014; Marsh, 2007; Pennamon, 2017; Ryalls, Benton, Li, & Barr, 2016; Spooren, Brockx, & Mortelmans, 2013; Theall, n.d.; Theall & Franklin, 2001), and (3) inaccuracy of the ratings, particularly in the context of online administrations.

In reviewing the validity studies of the 1970s and 1980s compared to the more recent wave of research findings, Nilson (2012) concluded that evidence substantiating the validity of student ratings had diminished significantly, which they didn't, especially in relation to achievement (Berk, 2016).

Their validity does not hinge on their relationship to student learning. She stated that their usefulness in decisions about faculty should be reexamined. Uttl et al's. (2017) meta-analysis of those multisection validity studies led the researchers to recommend that institutions should abandon student ratings as a measure of teaching effectiveness. The latest unflinching scrutiny of their analyses by Ryalls, Benton, and Li (2016) repudiated most of their conclusions.

Despite these attacks on the technical soundness of student ratings, they have garnered significant support. At present, a consensus of experts on student ratings agrees that properly-constructed scales, whether home-grown or commercial, used and interpreted appropriately, are far superior technically in their reliability and validity to all proposed alternative measures based on the vast psychometric research that has accumulated over several decades (Arreola, 2007; Benton & Cashin, 2012, 2014; Benton & Ryalls, 2016; Benton & Li, 2017; Berk, 2006, 2013c, 2013d, 2018; Hativa, 2014a, 2014b; Seldin, 2006; Theall, n.d.; Theall, Abrami, & Mets, 2001).

Misuse of scales and misinterpretation of ratings. Although guidelines, instructions, and manuals usually accompany student rating scales, they are still administered at many institutions under uncontrolled, unstandardized, and/or inappropriate conditions which can significantly decrease the response rate and render the answers invalid (Berk, 2006, 2013d). The numerous procedures available to maximize response rate for both face-to-face (f2f) and online administrations are often ignored (Berk, 2006, 2012, 2013a, 2013d). Even worse are the misinterpretations of the ratings for instructional changes and administrators' decisions about faculty (Boysen, Kelly, Raesly, & Casner, 2013; Linse, 2017). The latter are frequently based on the meaningless, trivial ranked differences in mean ratings (Berk, 2006, 2013d; Boysen, 2015a, 2015b) and the misuse of global ratings (Berk, 2013b). Both administrators and faculty need to be Mirandized on the proper use and interpretation of ratings for the decisions being made in their institution (see Berk, 2006; Linse, 2017).

Inadequate source of evidence for decision making. Based on these reported limitations and weaknesses, student ratings can provide only one portion of the information needed to infer teaching effectiveness. Unfortunately, that is pretty much all that is available at most colleges and universities. When those ratings alone are used for decision-making, the decisions will usually be based on incomplete and biased evidence.

Without additional evidence of teaching effectiveness, *student ratings can lead to incorrect, unfair, and evil career decisions about faculty* that can affect their contract renewal, annual salary increase, merit pay, professional development, and promotion and tenure (Wines & Lau, 2006). Administrators' pushing to use only student ratings for these decisions continues unabated. Even discriminatory practices based on age, gender, race, ethnicity, sexual orientation, religion, and other protected classes may occur, knowingly or unknowingly, because of how these scales are used (U.S. EEOC, 2010).

Conclusion. The preceding limitations, weaknesses, and critical issues related to student ratings should be considered in the process to evaluate teaching. Put most simply, s*tudent ratings from well-constructed scales are a necessary, but not sufficient, source of evidence to evaluate teaching comprehensively.*

If you or someone you know has serious reservations with the value assigned to student ratings in the last statement, you can now vent your outrage below:

> ***OUTRAGE VENT:*** Hold this box up to your mouth and express your outrage in the area in parentheses below in a loud and clear voice. Okay? Go.
>
> (Scream Your Outrage HERE)
>
> Okay. That's enough. You drooled on the box. That was unnecessary. You should be ashamed of yourself. Stop already. Calm down. Thank you for your input and saliva DNA sample.

Multiple Sources of Evidence

Since the 1990s, the practice of augmenting student ratings with other data sources of teaching effectiveness has been gaining traction in com-

munity colleges, liberal arts colleges, universities, medical schools/colleges (Berk, 2005, 2006; Canale, Herdklotz, & Wild, 2012; Tobin, Mandernach, & Taylor, 2015), and other institutions. 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; Benton & Li, 2017; Benton & Ryalls, 2016; Berk, 2005, 2006, 2013d, 2018; Berk, Naumann, & Appling, 2004; Braskamp & Ory, 1994; Cashin, 2003; Gravestock & Gregor-Greenleaf, 2008; Hoyt & Pallett, 1999; Knapper & Cranton, 2001; Ory, 2001; Seldin, 2006; Theall & Feldman, 2007; Theall & Franklin, 1990).

In fact, several comprehensive models of "faculty evaluation" that include multiple sources of evidence have been proposed (Arreola, 2007; Berk, 2006, 2009b, 2013d, 2018; Braskamp & Ory, 1994; Centra, 1993; Canale, Herdklotz, & Wild, 2012; Gravestock & Gregor-Greenleaf, 2008; Tobin et al., 2015). Some models attach greater weight to student and peer ratings and less weight to self, administrator, alumni ratings, and other sources. For online courses, Tobin et al. (2015) recommends also adding a variety of data analytics. All of these models can be used to arrive at formative and summative decisions.

15 Sources

What are the options? After scouring the literature, I dug up 15 potential sources of evidence of teaching effectiveness. These put student ratings into a broader context. The major categories of sources include students, instructor, other faculty, administrator, and employer. All of the sources apply to f2f, online, and blended/ hybrid courses. Here is a list of those sources:

STUDENTS

1. Student End-of-Course Ratings
2. Student Midterm Feedback
3. Student Exit and Alumni Ratings
4. Student Outcome Measures

INSTRUCTOR

5. Self-Ratings
6. Teaching Scholarship
7. Teaching Awards

OTHER FACULTY

8. Peer Classroom Observations,
9. Peer Review of Course Materials
10. External Expert Ratings
11. Mentor's Advice
12. Video Classroom Review
13. Teaching/Course Portfolio Review

ADMINISTRATOR

14. Administrator Ratings

EMPLOYER

15. Employer Ratings

Berk (2005, 2006, 2018) critically examined the value and contribution of these sources for measuring teaching effectiveness. The remainder of this article is a *CliffsNotes*® update and extension of the conclusions from those reviews based on the current state of research and practice. It provides practical guidelines to follow in the selection of specific sources of evidence for each type of decision and recommendations on how to proceed in your department.

How Do You Select the Right Source?

Triangulation. So far, what one simple conclusion can be drawn? "*This is really boring! Get to the point."* Wait! There must be some legitimate conclusion. "*Oh. There is no perfect source of evidence*." Bingo! Every source is different in form and substance from all of the other sources and can supply unique information. However, all sources are also fallible, usually in ways distinct from each other. For example, the unreliability and biases of peer observation ratings are not the same as those of student ratings; student ratings have other weaknesses (Marsh, 2007; Nilson, 2012).

So, what should you do? Since no single source of evidence can get the job done, draw from three or more different sources. The strengths of each source can 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, Naumann, & Berk, 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, online virtual class, or hybrid-time class, 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. The quality of the sources chosen should be beyond reproach.

Complementary multiple sources. At present, there is a paucity of empirical evidence to support the use of any particular combination of sources (e.g., Barnett, Matthews, & Jackson, 2003; Stalmeijer et al., 2010; Stehle, Spinath, & Kadmon, 2012). However, there is evidence on the relationships between student ratings and several other measures that supports their complementarity (*SYLLABLE ALERT:* 6 syllables? Are you kidding me?).

Benton and Cashin's (2012) research review reported the relationships between student ratings and ratings from observers, self, alumni, and administrators, which were low to moderate. Here are the actual validity coefficients with student ratings: trained external observers (.50 with global ratings), self (.30–.45), alumni (.54–.80)], administrators (.47–.62; .39 with global ratings). Those correlations indicate there are a lot of new information and insights on teaching to be gained by tapping those additional sources of evidence.

Beyond student ratings, is it worth the extra effort, time, cost, and aggravation to develop the additional measures mentioned previously? Are you squirming, queasy, or shuddering at the prospect of undertaking that initiative? Those side effects are normal. This is not for the faint of heart. Consider what you have to gain in the decisions you make.

As you build your instruments, it should become clear that they are intended to measure different teaching behaviors that contribute to teaching effectiveness. Each measure should cover a separate chunk of behaviors that are complementary, not redundant; however, some overlap of behaviors may be justified for corroboration. Every single tool should contribute significant, new information not measured by existing instruments to complete the puzzle that defines teaching effectiveness.

Matching Sources of Evidence to the Decision

Consider all 15 sources of evidence currently available based on my previous reviews and bottom line recommendations according to the research and experiences of others (Berk, 2005, 2006, 2009b, 2013d, 2018). *The decision should drive the choices of evidence*. Think carefully about the decision regarding the timeframe, conditions, information needed, and the faculty about whom the decision will be made. Which sources seem to be most appropriate for your decisions? Pick the highest quality sources for the specific decision. Prioritize the sources before you begin the task of collecting the evidence, which may involve the design and construction of new measures.

Currently, in addition to chair and dean ratings (68–79%), the most widely used sources for summative decisions in liberal arts colleges are student end-of-course ratings (94.2%), self-ratings (67.6%), peer classroom observation (60.4%), and peer review of course materials (52.5%) (Miller & Seldin, 2014). You can pick the pieces to assemble your collection of sources appropriate for your faculty. To jump-start your selection of sources, here are my top picks, based on the literature, for formative, summative, and program decisions:

Formative Decisions *(instructor improves and shapes the quality of teaching)*

- Student end-of-course ratings
- Student midterm feedback
- Peer classroom observation
- Peer review of course materials
- Self-ratings
- Video classroom review
- Mentor's advice
- External expert ratings

Summative Decisions *(administrator's annual review for contract renewal and merit pay)*

- Student end-of-course ratings
- Self-ratings
- Teaching scholarship
- Peer classroom observation (report written expressly for summative decision)
- Peer review of course materials (report written expressly for summative decision)
- Mentor's review (progress report written expressly for summative decision)

Summative Decisions *(committee's review for promotion and tenure)*

- Teaching/course portfolio review (across several years' courses)

Program Decisions *(faculty review of curriculum, admissions & graduation requirements, & program effectiveness)*

- Student end-of-course ratings
- Student exit and alumni ratings
- Student outcome measures

- Employer ratings

You probably noticed one particular source among the potential 15 that was conspicuously omitted from most of my recommended sources: learning outcome measures. Suffice it to say, for now, isolating students' course achievement at one point in time or their gains over time that are attributable directly to teaching is nearly impossible (Berk, 2014, 2016; Everson, 2017). The complexity increases considerably when attempting to compare faculty who teach different courses with different measures. It would be extremely difficult to defend student performance as a valid source of evidence of teaching effectiveness for any individual decision.

The multiple sources that were recommended previously for each decision can be configured into the *360º multisource feedback (MSF) model* of assessment (Berk, 2006, 2009a, 2009b) or another model for accreditation documentation of teaching evaluation. The sources for each decision may be added gradually to the model. Building the model is an ongoing process custom-tailored for each department or institution.

Final Recommendations for Practice

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

You're probably totally flummoxed by now. You may have some questions and certainly a few decisions to make. The first question may be "Where do I begin?" As you take the plunge into the process of designing a teaching evaluation program, I offer the following tips:

1. *Assemble a small faculty ad hoc committee.* Handpick appropriate "teachers" for your committee members, including at least one professor with expertise in measurement and evaluation. Add a couple students to provide their perspectives on the items they will be answering. Work will be involved.
2. *Map the outcomes for the semester (or quarter) and year*. Discuss a plan of attack. What are the highest priorities? Consider whether accreditation review is on the horizon or somewhere else. That could change the priorities.
3. *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 (Berk, 2010; Boysen, 2016). Decide what has to be done and who will do it.
4. *Review the other sources of evidence* with your faculty to decide the next steps. As you consider these sources in the gene pool of ideas, a few may prove fertile in your department. All stakeholders must be involved in these decisions. Don't be disheartened by the inevitable pushback.

   Just take this one step at a time. After you have spent a little time scratching your head and wondering what to do, decide which sources your faculty will embrace to reflect best practices in teaching? Weigh the pluses and minuses of the different sources. Prepare options for your faculty.
5. *Decide which combination of sources is best* for your faculty. Identify which sources should be used—although prepared differently—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.
6. *Design a detailed plan to build those sources*, one at a time, gradually, to create an evaluation model for each decision (see Berk, 2009b).

> ***Consultant Recommendation:*** If you're not sure how to proceed, talk to Farmers® Insurance. They know a thing or two because they've seen a thing or two. THEY ARE FARMERS®. Bum Bee Dee Bum, Bum Bum Bum. We now resume Recommendation 6 already in progress.

Delegate responsibility for and ownership of the various tasks involved. Faculty must make a professional commitment to "put it on the line," not just tip-toe near it. (*REMEMBER:* Administrators do not have time for these steps. They just need the data that faculty has agreed to use for decision making.)

Whatever combination of sources you choose to use, take the time and make an effort to design the scales, administer the scales, and report the results appropriately. Compared to what you are now using to make decisions, the new combination may turn ordinary sources into anything but. *The accuracy of faculty evaluation decisions depends on the integrity of the process and the validity and reliability of the multiple sources of evidence you collect.* Multiple sources are the uber-solution to evaluate teaching.

Tackling this endeavor may seem like a Sisyphean task (*GREEK FLASHBACK:* Sisyphus is remembered for pushing a Buick Regal up a mountain, only to have it roll back and smoosh him into a pancake). Like the Buick, you will probably receive pushback from some faculty, but keep in mind that you are not alone in this process. Your faculty and administrators are all vested. The measures that result should not have the earmarks of cobbled-together, made-by-committee products. The faculty's careers depend on those products. Solicit their input at every decision step of this journey.

You have the opportunity to make a brand new start of your faculty evaluation program. If you can make it in your institution, you can make it anywhere. Whatever package of sources your faculty produces, celebrate that accomplishment with appropriate pomp and circumstance or, at minimum, an upturned barrel of Gatorade®.

References

Addison, W. E., & Stowell, J. R. (2012). Conducting research on student evaluations of teaching. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 1–12). E-book retrieved from 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. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

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

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 handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system* (3rd ed.). Bolton, MA: Anker.

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.

Barre, E. (2015a, July 9). Do student evaluations of teaching really get an "F"? Houston, TX: Center for Teaching Excellence, Rice University. Retrieved from http://cte.rice.edu/blogarchive/2015/07/09/studentevaluations

Barre, E. (2015b, July 28). Academic blogging and student evaluation click bait: A follow-up. Houston, TX: Center for Teaching Excellence, Rice University. Retrieved from http://cte.rice.edu/blogarchive/2015/07/28/studentevaluationsfollowup

Basow, S. A., & Martin, J. L. (2012). Bias in student evaluations. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 40–49). E-book retrieved from http://teachpsych.org/ebooks/evals2012/index.php

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 from http://www.theideacenter.org/sites/default/files/idea-paper_50.pdf

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory & research* (Vol. 29, pp. 279–326). Dordrecht, The Netherlands: Springer.

Benton, S. L., & Li, D. (2017). *IDEA student ratings of instruction and RSVP* (*IDEA Paper No. 66*). Manhattan, KS: The IDEA Center. Retrieved from https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_66.pdf

Benton, S. L., & Ryalls, K. R. (2016). *Challenging misconceptions about student ratings of instruction* (*IDEA Paper No. 58*). Manhattan, KS: The IDEA Center.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48−62. Retrieved 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.

Berk, R. A. (2009a). Beyond student ratings: "A whole new world, a new fantastic point of view." *Essays on Teaching Excellence, 20*(1). Retrieved from 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. (2010). The secret to the "best" ratings from any evaluation scale. *Journal of Faculty Development, 24*(1), 37–39.

Berk, R. A. (2012). Top 20 strategies to increase the online response rates of student rating scales. *International Journal of Technology in Teaching and Learning, 8*(2), 98–107.

Berk, R. A. (2013a). Face-to-face versus online course evaluations: A "consumer's guide" to seven strategies. *Journal of Online Learning and Teaching, 9*(1), 140–148.

Berk, R. A. (2013b). Should global items on student rating scales be used for summative decisions? *Journal of Faculty Development, 27*(1), 57–61.

Berk, R. A. (2013c). Top 5 flashpoints in the assessment of teaching effectiveness. *Medical Teacher, 35*(1), 15–26. doi: 10.3109/0142159X.2012.732247

Berk, R. A. (2013d). *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.

Berk, R. A. (2014). Should student outcomes be used to evaluate teaching? *Journal of Faculty Development, 28*(2), 87–96.

Berk, R. A. (2016). Value of value-added models based on student outcomes to evaluate teaching. *Journal of Faculty Development, 30*(3), 73–81.

Berk, R. A. (2018). Beyond student ratings: 14 Other sources of evidence to evaluate teaching. In R. Ellis & E. Hogard (Eds.), *Handbook of quality assurance for university teaching.* London, UK: Routledge.

Berk, R. A., Naumann, P. L., & Appling, S. E. (2004). Beyond student ratings: Peer observation of classroom and clinical teaching. *International Journal of Nursing Education Scholarship, 1*(1), 1–26.

Berrett, D. (2015, November 29). Can the student course evaluation be redeemed? *The Chronicle of Higher Education*. Retrieved from http://www.chronicle.com/article/Can-the-Student-Course/234369?cid=rclink

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27-41. doi:10.1016/j.jpubeco.2016.11.006

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpenResearch*. Retrieved from https://www.scienceopen.com/document_file/25ff22be-8a1b-4c97-9d88-084c8d98187a/ScienceOpen/3507_XE6680747344554310733.pdf

Boysen, G. A. (2015a). Preventing the overinterpretation of small mean differences in student evaluations of teaching: An evaluation of warning effectiveness. *Scholarship of Teaching and Learning in Psychology, 1*(4), 269–282. Retrieved from http://dx.doi.org/10.1037/stl0000042

Boysen, G. A. (2015b). Uses and misuses of student evaluations of teaching: The interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology, 42*(2), 109–118. Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/0098628315569922

Boysen, G. A. (2016). Using student evaluations to improve teaching: Evidence-based recommendations. *Scholarship of Teaching and Learning in Psychology, 2*(4), 273–284. Retrieved from http://dx.doi.org/10.1037/stl0000069

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2013). The (mis) interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641–656. doi:10.1080/02602938.2013.860950

Braga, M., Paccagnella, M. & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71–88. Retrieved from https://doi.org/10.1016/j.econedurev.2014.04.002

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

Burt, S. (2015, May 15). Why not get rid of student evaluations? The answer requires us to think about power, *Slate*. Retrieved from http://www.slate.com/articles/life/education/2015/05/a_defense_of_student_evaluations_they_re_ biased_misleading_and_extremely.html

Calderon, T. G., Gabbin, A. L., & Green, B. P. (1996). *Report of the committee on promoting evaluating effective teaching*. Harrisonburg, VA: James Madison University.

Canale, A. M., Herdklotz, C., & Wild, L. (2012, November 13). *Evaluation of teaching effectiveness: Benchmark report and recommendations.* Rochester, NY: The Wallace Center at Rochester Institute of Technology, Office of Faculty Career Development. Retrieved from http://www.rit.edu/academicaffairs/facultydevelopment/sites/rit.edu.academicaffairs.facultydevelopment/files/docs/Evaluation_of_Teaching_Effectiveness.pdf

Cashin, W. E. (1989). *Defining and evaluating college teaching* (*IDEA Paper No. 21*). Manhattan, KS: The IDEA Center.

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.

Cohen P. A., & McKeachie, W. J. (1980). The role of colleagues in the evaluation of teaching. *Improving College and University Teaching, 28,* 147–154.

Coren, S. (2001). Are course evaluations a threat to academic freedom? In S. E. Kahn & D. Pavlich (Eds.), *Academic freedom and the inclusive university* (pp. 104–117). Vancouver: University of British Columbia Press.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198–1208.

Duckworth, A. (2016). *Grit: The power of passion and perseverance*. New York, NY: Scribner.

Everson, K. C. (2017). Value-added modeling and educational accountability: Are we answering the real questions? *Review of Educational Research, 87*(1), 35–70. Retrieved from http://journals.sagepub.com/doi/full/10.3102/0034654316637199

Freyd, M. (1923). A graphic rating scale for teachers. *Journal of Educational Research*, *8*(5), 433–439.

Gooblar, D. (2017, May 31). No, student evaluations aren't "worthless." *Chronicle Vitae*. Retrieved from https://chroniclevitae.com/news/1814-no-student-evaluations-aren-t-worthless?cid=at&utm_source=at&utm_medium=en&elqTrackId=82121c92a9514c93bc15c5b7601dc6f9&elq=7e78f00e5df94d5280e706170379a9c5&elqaid=14140&elqat=1&elqCampaignId=5932

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

Green, B. P., Calderon, T. G., & Reider, B. P. (1998). A content analysis of teaching evaluation instruments used in accounting departments. *Issues in Accounting Education*, *13*(1), 15–30.

Griffin, A., & Cook, V. (2009). Acting on evaluation: Twelve tips from a national conference on student evaluations. *Medical Teacher, 31*, 101–104.

Hativa, N. (2014a). *Student ratings of instruction: A practical approach to designing, operating, and reporting*. Oron Publications. nhativa@gmail.com

Hativa, N. (2014b). *Student ratings of instruction: Recognizing effective teaching*. Oron Publications. nhativa@gmail.com

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.

Kamenetz, A. (2014, September 26). Student course evaluations get an 'F'. *NPR Ed: How Learning Happens.* Retrieved from http://www.npr.org/sections/ed/2014/09/26/345515451/student-course-evaluations-get-an-f

Keig, L. W., & Waggoner, M. D. (1994). *Collaborative peer review: The role of faculty in improving college teaching* (*ASHE/ERIC Higher Education Report, No. 2*). Washington, DC: Association for the Study of Higher Education.

Kite, M. E. (Ed.). (2012). *Effective evaluation of teaching: A guide for faculty and administrators*. E-book retrieved from the Society for the Teaching of Psychology website http://teachpsych.org/ebooks/evals2012/index.php

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.

Li, D., & Benton, S. L. (2017). *The effects of instructor gender and discipline group on student ratings of instruction* (*IDEA Research Report No. 10*). Manhattan, KS: The IDEA Center.

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation, 54,* 94–106. Retrieved from http://www.schreyerinstitute.psu.edu/pdf/SRTE_Guidelines_Linse_JSEE_2017.pdf

Macnell, L., Driscoll, A., & Hunt, A. N. (2014). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303. doi:10.1007/s10755-014-9313-4

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, The Netherlands: Springer.

Miller, J. E., & Seldin, P. (2014, May-June). Changing practices in faculty evaluation: Can better evaluation make a difference? *Bulletin of the AAUP,100*(3). Retrieved from www.aaup.org/article/changing-practices-faculty-evaluation#.WNawXW_yvIV

Nilson, L. B. (2012). Time to raise questions about student ratings. In J. E. Groccia & L. Cruz (Eds.), *To improve the academy: Resources for*

faculty, instructional, and organizational development (Vol. 31, pp. 213–228). San Francisco: Jossey-Bass.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. In K. G. Lewis (Ed.), *Techniques and strategies for interpreting student evaluations* (Special issue) (*New Directions for Teaching and Learning, No. 87*) (pp. 3–15). San Francisco: Jossey-Bass.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* (Special issue) (*New Directions for Institutional Research, No. 109*) (pp. 27–44). San Francisco: Jossey-Bass.

Pennamon, T. (2017, June 22). Student evaluations at center of American University tenure fight. *Diverse: Issues in Higher Education*. Retrieved from http://diverseeducation.com/article/98161/?utm_campaign=DIV1706%20DAILY%20NEWSLETTER%20JUN23&utm_medium=email&utm_source=Eloqua

Quintana, C. (2017, May 30). As summer sets in, a chance to regard the good, bad, and ugly of student evaluations. *The Chronicle of Higher Education*. Retrieved from http://www.chronicle.com/article/As-Summer-Sets-In-a-Chance-to/240203?cid=db&elqTrackId=f40f875f469340448ee2b8647b354cc5&elq=32c4887aa10d4d4885390a178bb44c01&elqaid=14197&elqat=1&elqCampaignId=5962

Ryalls, K. R., Benton, S. L., & Li, D. (2016, November). *Response to "Zero correlation between evaluations and learning"* (*IDEA Editorial Note #3*). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development. Retrieved from http://www.ideaedu.org/Portals/0/Uploads/Documents/Response_to_Zero_Correlation_Between_Evaluations_Teaching.pdf

Ryalls, K. R., Benton, S. L., Li, D., & Barr J. (2016). *Response to "Bias against female instructors"* (*IDEA Editorial Note*). Manhattan, KS: The IDEA Center. Retrieved from http://ideaedu.org/research-and-papers/editorial-notes/response-to-bias-against-female-instructors/

Schuman, R. (2014, April 24). Needs improvement: Student evaluations of professors aren't just biased and absurd—they don't even work. *Slate.* Retrieved from http://www.slate.com/articles/life/education/2014/04/student_evaluations_of_college_professors_are_biased_and_worthless.html

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.

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*(3), 1–45.

Sproule, R., & Valsan, C. (2009). The student evaluation of teaching: Its failure as a research program, and as an administrative guide. *Economic Interferences, 11*(25), 125–150. Retrieved from http://www.amfiteatrueconomic.ro/temp/Article_641.pdf

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.

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, 53*(8), 888–904. doi: 10.1007/s11162-012-9260-9

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. Retrieved from http://www.hea.ie/files/files/DES_Higher_Ed_Main_Report.pdf.

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science, 11*(6), 800–816. doi:10.1177/1745691616650284

Surgenor, P. W. G. (2011). Obstacles and opportunities: Addressing the growing pains of summative student evaluation of teaching. *Assessment & Evaluation in Higher Education*, 1–14, iFirst Article. doi: 10.1080/02602938.2011.635247

Svinicki, M., & McKeachie, W. J. (2014). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (14th ed.). Belmont, CA: Wadsworth.

Theall, M. (n.d.). Student ratings: Myths vs. research evidence. Paper prepared for the Teaching and Learning Center, Brigham Young University. Retrieved from http://facultyaffairs.arizona.edu/sites/facultyaffairs/files/student-ratings.pdf

Theall, M., Abrami, P. C., & Mets, L. A. (Eds.) (2001). *The student ratings debate: Are they valid*? *How can we best use them*? (*New Directions for Institutional Research, No. 109*). San Francisco: Jossey-Bass.

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.

Theall, M., & Franklin, J. L. (1990). Student ratings in the context of complex evaluation systems. In M. Theall & J. L. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (*New Directions for Teaching and Learning, No. 43*) (pp. 17–34). San Francisco: Jossey-Bass.

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

Tobin, T. J., Mandernach, B. J., & Taylor, A. H. (2015). *Evaluating online teaching: Implementing best practices*. San Francisco: Jossey-Bass.

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

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54,* 22–42. Retrieved from https://www.chrisstucchio.com/blog_media/2016/assorted_links_nov_3_2016/Meta_analysis_of_faculty's_teaching_effectiveness_Student_evaluation_of_teaching_ratings_and_student_learning_are_not_related__student_evaluations_meta_analysis.pdf

Wieman, C. (2015, February 6). A better way to evaluate undergraduate teaching. *Change: The Magazine of Higher Learning, 47*(1). Retrieved from http://www.tandfonline.com/doi/full/10.1080/00091383.2015.996077

Wines, W. A., & Lau, T. J. (2006). Observations on the folly of using student evaluations of college teaching for faculty evaluation, pay, and retention decisions and its implications for academic freedom. *William & Mary Journal of Women and the Law*, *13*(1), 167–202.

Ronald A. Berk, PhD, is professor emeritus, biostatistics and measurement, and former assistant dean for teaching at The Johns Hopkins University. Now he is a 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.