An Evaluation of Course Evaluations

Published in *ScienceOpen*:
https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015-dc99b467b6e4?0

DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Philip B. Stark*

Department of Statistics, University of California, Berkeley
Berkeley, CA 94720, United States
stark@stat.berkeley.edu

Richard Freishtat

Center for Teaching and Learning, University of California, Berkeley
Berkeley, CA 94720, United States
rfreishtat@berkeley.edu

26 September 2014

* Corresponding author. E-mail: stark@stat.berkeley.edu

26 September 2014

Student ratings of teaching have been used, studied, and debated for almost a century. This article examines student ratings of teaching from a statistical perspective. The common practice of relying on averages of student teaching evaluation scores as the primary measure of teaching effectiveness for promotion and tenure decisions should be abandoned for substantive and statistical reasons: There is strong evidence that student responses to questions of "effectiveness" do not measure teaching effectiveness. Response rates and response variability matter. And comparing averages of categorical responses, even if the categories are represented by numbers, makes little sense. Student ratings of teaching are valuable when they ask the right questions, report response rates and score distributions, and are balanced by a variety of other sources and methods to evaluate teaching.

Since 1975, course evaluations at *University of California, Berkeley* have asked:

> Considering both the limitations and possibilities of the subject matter and course, how would you rate the overall teaching effectiveness of this instructor?
>
> 1 (not at all effective), 2, 3, 4 (moderately effective), 5, 6, 7 (extremely effective)

Among faculty, student evaluations of teaching (SET) are a source of pride and satisfaction—and frustration and anxiety. High-stakes decisions including tenure and promotions rely on SET. Yet it is widely believed that they are primarily a popularity contest; that it's easy to "game" ratings; that good teachers get bad ratings and *vice versa*; and that rating anxiety stifles pedagogical innovation and encourages faculty to water down course content. What's the truth?

We review statistical issues in analyzing and comparing SET scores, problems defining and measuring teaching effectiveness, and pernicious distortions that result from using SET scores as a proxy for teaching quality and effectiveness. We argue here--and the literature shows--that students are in a good position to evaluate *some* aspects of teaching, but SET are at best tenuously connected to teaching effectiveness (Defining and measuring teaching effectiveness are knotty problems in themselves; we discuss this below). Other ways of evaluating teaching can be combined with student comments to produce a more reliable and meaningful composite. We make recommendations regarding the use of SET and discuss new policies implemented at *University of California, Berkeley*, in 2013.

Background

SET scores are the most common method to evaluate teaching (Cashin, 1999; Clayson, 2009; Davis, 2009; Seldin, 1999). They define "effective teaching" for many purposes. They are popular partly because the measurement is easy and takes little class or faculty time. Averages of SET ratings have an air of objectivity simply by virtue of being numerical. And comparing an instructor's average rating to departmental averages is simple. However, questions about using SET as the sole source of evidence about teaching for merit and promotion, and the efficacy of evaluation questions and methods of interpretation persist (Pounder, 2007).

Statistics and SET

Who responds?

Some students do not fill out SET surveys. The *response rate* will be less than 100%. The lower the response rate, the less representative the responses might be: there's no reason nonresponders should be like responders--and good reasons they might not be. For instance, anger motivates people to action more than satisfaction does. Have you ever seen a public demonstration where people screamed "we're content!"? (See, e.g., http://xkcd.com/470/)

Nonresponse produces uncertainty: Suppose half the class responds, and that they rate the instructor's handwriting legibility as 2. The average for the entire class might be as low as 1.5, if all the "nonresponders" would also have rated it 1. Or it might be as high as 4.5, if the nonresponders would have rated it 7.

Some schools require faculty to explain low response rates. This seems to presume that it is the instructor's fault if the response rate is low, and that a low response rate is in itself a sign of bad teaching. Consider these scenarios:

(1) The instructor has invested an enormous amount of effort in providing the material in several forms, including online materials, online self-test exercises, and webcast lectures; the course is at 8am. We might expect attendance and response rates to in-class evaluations to be low.

(2) The instructor is not following any text and has not provided notes or

supplementary materials. Attending lecture is the only way to know what is covered. We might expect attendance and response rates to in-class evaluations to be high.

(3) The instructor is exceptionally entertaining, gives "hints" in lecture about exams; the course is at 11am. We might expect high attendance and high response rates for in-class evaluations.

The point: Response rates themselves say little about teaching effectiveness. In reality, if the response rate is low, the data should not be considered representative of the class as a whole. An explanation solves nothing.

Averages of small samples are more susceptible to "the luck of the draw" than averages of larger samples. This can make SET in small classes more extreme than evaluations in larger classes, even if the response rate is 100%. And students in small classes might imagine their anonymity to be more tenuous, perhaps reducing their willingness to respond truthfully or to respond at all.

Averages

Personnel reviews routinely compare instructors' average scores to departmental averages. Such comparisons make no sense, as a matter of Statistics. They presume that the difference between 3 and 4 means the same thing as the difference between 6 and 7. They presume that the difference between 3 and 4 means the same thing to different students. They presume that 5 means the same thing to different students and to students in different courses.

They presume that a 3 “balances” a 7 to make two 5s. For teaching evaluations, there’s no reason any of those things should be true (See, e.g., McCullough & Radson, 2011).

SET scores are *ordinal categorical* variables: The ratings fall in categories that have a natural order, from worst (1) to best (7). But the numbers are *labels*, not *values*. We could replace the numbers with descriptions and no information would be lost: The ratings might as well be “not at all effective,” … , “extremely effective.” It doesn’t make sense to average labels. Relying on averages equates two ratings of 5 with ratings of 3 and 7, since both sets average to 5.

They are not equivalent, as this joke shows: Three statisticians go hunting. They spot a deer. The first statistician shoots; the shot passes a yard to the left of the deer. The second shoots; the shot passes a yard to the right of the deer. The third one yells, “we got it!”

Scatter matters

Comparing an individual instructor’s average with the average for a course or a department is meaningless: Suppose that the departmental average for a particular course is 4.5, and the average for a particular instructor in a particular semester is 4.2. The instructor’s rating is below average. How bad is that?

If other instructors get an average of exactly 4.5 when they teach the course, 4.2 might be atypically low. On the other hand, if other instructors get 6s half the time and 3s half the time, 4.2 is well within the spread of scores. Even if

averaging made sense, the mere fact that one instructor's average rating is above or below the departmental average says little. We should report the *distribution* of scores for instructors and for courses: the percentage of ratings in each category (1–7). The distribution is easy to convey using a bar chart.

All the children are above average

At least half the faculty in any department will have average scores at or below median for that department. Deans and Chairs sometimes argue that a faculty member with below-average teaching evaluations is an excellent teacher—just not as good as the other, superlative teachers in that department. With apologies to Garrison Keillor, all faculty members in all departments cannot be above average.

Comparing incommensurables

Students' interest in courses varies by course type (e.g., prerequisite versus major elective). The nature of the interaction between students and faculty varies with the type and size of courses. Freshmen have less experience than seniors. These variations are large and may be confounded with SET (Cranton & Smith, 1986; Feldman, 1984, 1978). It is not clear how to make fair comparisons of SET across seminars, studios, labs, prerequisites, large lower-division courses, required major courses, etc (See, e.g., McKeachie, 1997).

Student Comments

Students are ideally situated to comment *about their experience* of the

course, including factors that influence teaching effectiveness, such as the instructor's audibility, legibility, and perhaps the instructor's availability outside class. They can comment on whether they feel more excited about the subject after taking the class, and—for electives—whether the course inspired them to take a follow-up course. They might be able to judge clarity, but clarity may be confounded with the difficulty of the material. While some student comments are informative, one must be quite careful interpreting the comments: faculty and students use the same vocabulary quite differently, ascribing quite different meanings to words such as "fair," "professional," "organized," "challenging," and "respectful" (Lauer, 2012). Moreover, it is not easy to compare comments across disciplines (Cashin, 1990; Cashin & Clegg, 1987; Cranton & Smith, 1986; Feldman, 1978), because the depth and quality of students' comments vary widely by discipline. In context, these comments are all glowing:

Physical Sciences class.

> "Lectures are well organized and clear"
>
> "Very clear, organized and easy to work with"

Humanities class.

> "Before this course I had only read two plays because they were required in High School. My only expectation was to become more familiar with

the works. I did not expect to enjoy the selected texts as much as I did, once they were explained and analyzed in class. It was fascinating to see texts that the author's were influenced by; I had no idea that such a web of influence in Literature existed. I wish I could be more 'helpful' in this evaluation, but I cannot. I would not change a single thing about this course. I looked forward to coming to class everyday. I looked forward to doing the reading for this class. I only wish that it was a year long course so that I could be around the material, graduate instructor's and professor for another semester."

What SET Measure

If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference.

-D. Huff (1954)

This is what we do with SET. We don't measure teaching effectiveness. We measure what students say, and pretend it's the same thing. We calculate statistics, report numbers, and call it a day.

What is effective teaching? One definition is that an effective teacher is skillful at creating conditions conducive to learning. Some learning happens no matter what the instructor does. Some students do not learn much no matter what the instructor does. How can we tell how much the instructor helped or hindered?

Measuring learning is hard: Grades are poor proxies, because courses and exams can be easy or hard (Beleche, Fairris and Marks, 2012). If exams were set by someone other than the instructor—as they are in some universities—we might be able to use exam scores to measure learning (See, e.g., http://xkcd.com/135/). But that's not how most universities work, and teaching to the test could be confounded with learning.

Performance in follow-on courses and career success may be better measures, but those measurements are hard to make. And how much of someone's career success can be attributed to a given course, years later?

There is a large research literature on SET, most of which addresses *reliability*: Do different students give the same instructor similar marks (See, e.g., Abrami, et al., 2001; Braskamp and Ory, 1994; Centra, 2003; Ory, 2001; Wachtel, 1998; Marsh and Roche, 1997)? Would a student rate the same instructor consistently later (See, e.g., Braskamp and Ory, 1994; Centra, 1993; Marsh, 2007; Marsh and Dunkin, 1992; Overall and Marsh, 1980)? That has nothing to do with

whether SET measure effectiveness. A hundred bathroom scales might all report your weight to be the same. That doesn't mean the readings are accurate measures of your *height*—or even your weight, for that matter.

Moreover, inter-rater reliability is an odd thing to worry about, in part because it's easy to report the full distribution of student ratings, as advocated above. Scatter matters, and it can be measured *in situ* in every course.

Observation versus Randomization

Most of the research on SET is based on *observational studies*, not *experiments*. In the entire history of Science, there are few observational studies that justify inferences about causes (A notable exception is John Snow's research on the cause of cholera; his study amounts to a "natural experiment." See http://www.stat.berkeley.edu/~stark/SticiGui/Text/experiments.htm#cholera for a discussion). In general, to infer causes, such as whether good teaching results in good evaluation scores, requires a *controlled, randomized experiment*: individuals are assigned to groups at random; the groups get different *treatments*; the outcomes are compared statistically across groups to test whether the treatments have different effects and to estimate the sizes of those differences.

Randomized experiments use a blind, non-discretionary chance mechanism to assign treatments to individuals. Randomization tends to mix individuals across groups in a balanced way. Absent randomization, other things can *confound* the effect of the treatment (See, e.g., http://xkcd.com/552/).

For instance, suppose some students choose classes by finding the professor reputed to be the most lenient grader. Such students might then rate that professor highly for an "easy A." If those students choose sequel courses the same way, they may get good grades in those easy classes too, "proving" that the first ratings were justified.

The best way to reduce confounding is to assign students randomly to classes. That tends to mix students with different abilities and from easy and hard sections of the prequel across sections of sequels. This experiment has been done at the U.S. Air Force Academy (Carrell and West, 2008) and Bocconi University in Milan, Italy (Braga, Paccagnella, and Pellizzari, 2011).

These experiments found that teaching effectiveness, as measured by subsequent performance and career success, is *negatively* associated with SET scores. While these two student populations might not be representative of all students, the studies are the best we have seen. And their findings are concordant.

What do student teaching evaluations measure?

SET may be *reliable*, in the sense that students often agree (Braskamp and Ory, 1994; Centra, 1993; Marsh, 2007; Marsh and Dunkin, 1992; Overall and Marsh, 1980). But that's an odd focus. We don't expect instructors to be equally effective with students with different background, preparation, skill, disposition, maturity, and "learning style." Hence, if ratings are extremely consistent, they probably don't measure teaching effectiveness: If a laboratory instrument always

gives the same reading when its inputs vary substantially, it's probably broken. There is no consensus on what SET do measure:

- SET scores are highly correlated with students' grade expectations (Marsh and Cooper, 1980; Short et al., 2012; Worthington, 2002)
- SET scores and enjoyment scores are related (In the UC Berkeley Department of Statistics in fall 2012, for the 1486 students who rated the instructor's overall effectiveness and their enjoyment of the course, the correlation between instructor effectiveness and course enjoyment was 0.75, and the correlation between course effectiveness and course enjoyment was 0.8.)
- SET can be predicted from the students' reaction to 30 seconds of silent video of the instructor; physical attractiveness matters (Ambady and Rosenthal, 1993).
- gender, ethnicity, and the instructor's age matter (Anderson and Miller, 1997; Basow, 1995; Cramer and Alexitch, 2000; Marsh and Dunkin, 1992; Wachtel, 1998; Weinberg et al., 2007; Worthington, 2002).
- omnibus questions about curriculum design, effectiveness, etc. appear most influenced by factors unrelated to learning (Worthington, 2002)

<u>What good are SET?</u>

Students are in a good position to observe some aspects of teaching, such as clarity, pace, legibility, audibility, and their own excitement (or boredom).

SET can measure these things; the statistical issues raised above still matter, as do differences between how students and faculty use the same words (Lauer, 2012).

But students cannot rate effectiveness--regardless of their intentions. Calling SET a measure of effectiveness does not make it one, any more than you can make a bathroom scale measure height by relabeling its dial "height." Averaging "height" measurements made with 100 different scales would not help.

What's better?

Let's drop the pretense. We will never be able to measure teaching effectiveness reliably and routinely. In some disciplines, measurement is possible but would require structural changes, randomization, and years of follow-up.

If we want to assess and improve teaching, we have to pay attention to the teaching, not the average of a list of student-reported numbers with a troubled and tenuous relationship to teaching. Instead, we can watch each other teach and talk to each other about teaching. We can look at student comments. We can look at materials created to design, redesign, and teach courses, such as syllabi, lecture notes, websites, textbooks, software, videos, assignments, and exams. We can look at faculty teaching statements. We can look at samples of student work. We can survey former students, advisees, and graduate instructors. We can look at the job placement success of former graduate students. Etc.

We can ask: Is the teacher putting in appropriate effort? Is she following

practices found to work in the discipline? Is she available to students? Is she creating new materials, new courses, or new pedagogical approaches? Is she revising, refreshing, and reworking existing courses? Is she helping keep the curriculum in the department up to date? Is she trying to improve? Is she supervising undergraduates for research, internships, and honors theses? Is she advising graduate students? Is she serving on qualifying exams and thesis committees? Do her students do well when they graduate?

Or, is she "checked out"? Does she use lecture notes she inherited two decades ago the first time she taught the course? Does she mumble, facing the board, scribbling illegibly? Do her actions and demeanor discourage students from asking questions? Is she unavailable to students outside of class? Does she cancel class frequently? Does she return student work with helpful comments? Does she refuse to serve on qualifying exams or dissertation committees?

In 2013, the University of California, Berkeley Department of Statistics adopted as standard practice a more holistic assessment of teaching. Every candidate is asked to produce a teaching portfolio for personnel reviews, consisting of a teaching statement, syllabi, notes, websites, assignments, exams, videos, statements on mentoring, or any other materials the candidate feels are relevant. The chair and promotion committee read and comment on the portfolio in the review. At least before every "milestone" review (mid-career, tenure, full, step VI), a faculty member attends at least one of the candidate's lectures and

comments on it, in writing. These observations complement the portfolio and student comments. Distributions of SET scores are reported, along with response rates. Averages of scores are not reported.

Classroom observation took the reviewer about four hours, including the observation time itself. The process included conversations between the candidate and the observer, the opportunity for the candidate to respond to the written comments, and a provision for a "no-fault do-over" at the candidate's sole discretion. The candidates and the reviewer reported that the process was valuable and interesting. Based on this experience, the Dean of the Division now recommends peer observation prior to milestone reviews.

Observing more than one class session and more than one course would be better. Adding informal classroom observation and discussion between reviews would be better. Periodic surveys of former students, advisees, and teaching assistants would bring another, complementary source of information about teaching. But we feel that using teaching portfolios and even a little classroom observation improves on SET alone.

The following sample letter is a redacted amalgam of chair's letters submitted with merit and promotion cases since the Department of Statistics adopted a policy of more comprehensive assessment of teaching, including peer observation:

Smith is, by all accounts, an excellent teacher, as confirmed by the

classroom observations of Professor Jones, who calls out Smith's ability to explain key concepts in a broad variety of ways, to hold the attention of the class throughout a 90-minute session, to use both the board and slides effectively, and to engage a large class in discussion. Prof. Jones's peer observation report is included in the case materials; conversations with Jones confirm that the report is Jones's candid opinion: Jones was impressed, and commented in particular on Smith's rapport with the class, Smith's sensitivity to the mood in the room and whether students were following the presentation, Smith's facility in blending derivations on the board with projected computer simulations to illustrate the mathematics, and Smith's ability to construct alternative explanations and illustrations of difficult concepts when students did not follow the first exposition.

While interpreting "effectiveness" scores is problematic, Smith's teaching evaluation scores are consistently high: in courses with a response rate of 80% or above, less than 1% of students rate Smith below a 6.

Smith's classroom skills are evidenced by student comments in teaching evaluations and by the teaching materials in her portfolio.

Examples of comments on Smith's teaching include:

> *I was dreading taking a statistics course, but after this class, I decided to major in statistics.*
>
> *the best I've ever met...hands down best teacher I've had in 10 years of university education*
>
> *overall amazing…she is the best teacher I have ever had*
>
> *absolutely love it*
>
> *loves to teach, humble, always helpful*
>
> *extremely clear … amazing professor*
>
> *awesome, clear*
>
> *highly recommended*
>
> *just an amazing lecturer*
>
> *great teacher … best instructor to date*
>
> *inspiring and an excellent role model*
>
> *the professor is GREAT*

Critical student comments primarily concerned the difficulty of the material or the homework. None of the critical comments reflected on the pedagogy or teaching effectiveness, only the workload.

I reviewed Smith's syllabus, assignments, exams, lecture notes, and other materials for Statistics X (a prerequisite for many majors), Y (a seminar course she developed), Z (a graduate course she developed for the revised MA program, which she has spearheaded), and Q (a topics course in her research area). They are very high quality and clearly the result of considerable thought and effort.

In particular, Smith devoted an enormous amount of time to developing online materials for X over the last five years. The materials required designing and creating a substantial amount of supporting technology, representing at least 500 hours per year of effort to build and maintain. The undertaking is highly creative and advanced the state of the art. Not only are those online materials superb, they are having an impact on pedagogy elsewhere: a Google search shows over 1,200 links to those materials, of which more than half are from other countries. I am quite impressed with the pedagogy, novelty, and functionality. I have a few minor suggestions about the content, which I will discuss with Smith, but those are a matter of taste, not of correctness.

The materials for X and Y are extremely polished. Notably, Smith assigned a term project in an introductory course, harnessing the power of inquiry-based learning. I reviewed a handful of the term projects, which were ambitious and impressive. The materials for Z and Q are also well organized and interesting, and demand an impressively high level of performance from the students. The materials for Q include a great selection of data sets and computational examples that are documented well. Overall, the materials are exemplary; I would estimate that they represent well over 1,500 hours of development during the review period.

Smith's lectures in X were webcast in fall, 2013. I watched portions of a dozen of Smith's recorded lectures for X—a course I have taught many times. Smith's lectures are excellent: clear, correct, engaging, interactive, well paced, and with well organized and legible boardwork. Smith does an excellent job keeping the students involved in discussion, even in large (300+ student) lectures. Smith is particularly good at keeping the students thinking during the lecture and of inviting questions and comments. Smith responds generously and sensitively to questions, and is tuned in well to the mood of the class.

Notably, some of Smith's lecture videos have been viewed nearly 300,000 times! This is a testament to the quality of Smith's pedagogy and reach. Moreover, these recorded lectures increase the visibility of the Department and the University, and have garnered unsolicited effusive thanks and praise from across the world.

Conversations with teaching assistants indicate that Smith spent a considerable amount of time mentoring them, including weekly meetings and observing their classes several times each semester. She also played a leading role in revising the PhD curriculum in the department.

Smith has been quite active as an advisor to graduate students. In addition to serving as a member of sixteen exam committees and more than a dozen MA and PhD committees, she advised three PhD recipients (all of whom got jobs in top-ten departments), co-advised two others, and is currently advising three more. Smith advised two MA recipients who went to jobs in industry, co-advised another who went to a job in government, advised one who changed advisors. Smith is currently advising a fifth. Smith supervised three undergraduate honors theses and two undergraduate internships during the review period.

This is an exceptionally strong record of teaching and mentoring for an assistant professor. Prof. Smith's teaching greatly exceeds expectations.

We feel that a review along these lines would better reflect whether faculty are dedicated teachers, the effort they devote, and the effectiveness their teaching; would comprise a much fairer assessment; and would put more appropriate attention on teaching.

Recap

- SET does not measure teaching effectiveness.
- Controlled, randomized experiments find that SET ratings are negatively associated with direct measures of effectiveness. SET seem to be influenced by the gender, ethnicity, and attractiveness of the instructor.
- Summary items such as "overall effectiveness" seem most influenced by

irrelevant factors.

- Student comments contain valuable information about students' *experiences*.
- Survey response rates matter. Low response rates make it impossible to generalize reliably from the respondents to the whole class.
- It is practical and valuable to have faculty observe each other's classes.
- It is practical and valuable to create and review teaching portfolios.
- Teaching is unlikely to improve without serious, regular attention.

Recommendations

1. Drop omnibus items about "overall teaching effectiveness" and "value of the course" from teaching evaluations: They are misleading.
2. Do not average or compare averages of SET scores: Such averages do not make sense statistically. Instead, report the distribution of scores, the number of responders, and the response rate.
3. When response rates are low, extrapolating from responders to the whole class is unreliable.
4. Pay attention to student comments—but understand their limitations. Students typically are not well situated to evaluate pedagogy.
5. Avoid comparing teaching in courses of different types, levels, sizes, functions, or disciplines.
6. Use teaching portfolios as part of the review process.

7. Use classroom observation as part of milestone reviews.
8. To improve teaching and evaluate teaching fairly and honestly, spend more time observing the teaching and looking at teaching materials.

References

Abrami, P.C., Marilyn, H.M. & Raiszadeh, F. (2001). Business students' perceptions of faculty evaluations. *The International Journal of Educational Management, 15*(1), 12–22.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431.

Anderson, K., & Miller, E.D. (1997). Gender and student evaluations of teaching. *PS: Political Science and Politics, 30*(2), 216-219.

Basow, S.A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*(4), 656-665.

Beleche, T., Fairris, D., & Marks, M. (2012). Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. *Economics of Education Review*, *31*(5), 709-719.

Braga, M., Paccagnella, M., & Pellizzari, M. (2011). Evaluating students' evaluations of professors. *Bank of Italy Temi di Discussione (Working Paper) No*, *825*.

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

Carrell, S.E., & West, J.E. (2008). *Does professor quality matter? Evidence from random assignment of students to professors* (No. w14081). National

Bureau of Economic Research.

Cashin, W.E. (1990). Students do rate different academic fields differently. In M. Theall and J. Franklin (eds.), *Student Ratings of Instruction: Issues for improving practice*. San Francisco: Jossey-Bass Inc.

Cashin, W.E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (ed.), *Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions.* Bolton, MA: Anker.

Cashin, W.E. and Clegg, V.L. (1987). *Are student ratings of different academic fields different?* Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.

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

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less coursework? *Research in Higher Education*, *44*(5), 495-518.

Clayson, D.E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16-30.

Cramer, K.M. & Alexitch, L.R. (2000). Student evaluations of college professors: identifying sources of bias. *Canadian Journal of Higher Education, 30*(2),

143-64.

Cranton, P.A. and Smith, R.A. (1986). A new look at the effect of course characteristics on student ratings of instruction. *American Educational Research Journal, 23*(1), 117–128.

Davis, B.G. (2009). *Tools for Teaching, 2nd edition*. San Francisco, CA: John Wiley & Sons.

Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't know. *Research in Higher Education, 9*, 199–242.

Feldman, K.A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*(11), 45–116.

Huff, D. (1954). *How To Lie With Statistics*, New York: W.W. Norton.

Lauer, C. (2012). A Comparison of Faculty and Student Perspectives on Course Evaluation Terminology. In *To Improve the Academy: Resources for Faculty, Instructional, and Organizational Development*, edited by J. Groccia & L. Cruz, 195-212. San Francisco, CA: Wiley & Sons, Inc.

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,* 319–383. Dordrecht, The Netherlands: Springer.

Marsh, H.W., & Cooper, T. (1980) *Prior subject interest, students evaluations, and instructional effectiveness* Paper presented at the annual meeting of the American Educational Research Association.

Marsh, H.W., & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research*, Vol. 8. New York: Agathon Press.

Marsh, H.W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187–1197.

McCullough, B. D., & Radson, D. (2011). Analysing student evaluations of teaching: Comparing means and proportions. *Evaluation & Research in Education*, *24*(3), 183–202.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225.

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, 87*, 3–15.

Overall, J.U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*, 321–325.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile?: An

analytical framework for answering the question. *Quality Assurance in Education*, *15*(2), 178-191.

Seldin, P. (1999). Building successful teaching evaluation programs. In P. Seldin (ed.), *Current practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions.* Bolton, MA: Anker.

Short, H., Boyle, R., Braithwaite, R., Brookes, M., Mustard, J., & Saundage, D. (2008). A comparison of student evaluation of teaching with student performance. In *OZCOTS 2008: Proceedings of the 6th Australian Conference on Teaching Statistics* (pp. 1–10).

Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191–211.

Weinberg, B.A., Fleisher, B.M., & Hashimoto, M. (2007). *Evaluating methods for evaluating instruction: The case of higher education (NBER Working Paper No. 12844)*. Retrieved 5 August 2013 from http://www.nber.org/papers/w12844http://www.nber.org/papers/w12844

Worthington, A.C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education, *Assessment and Evaluation in Higher Education*, *27*(1), 49–64.