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

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)

What good are SET?

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.

Review of Educational Research
December 2013, Vol. 83, No. 4, pp. 598–642
DOI: 10.3102/0034654313496870

On the Validity of Student Evaluation of Teaching: The State of the Art

Pieter Spooren, Bert Brockx, and Dimitri Mortelmans
University of Antwerp

This article provides an extensive overview of the recent literature on student evaluation of teaching (SET) in higher education. The review is based on the SET meta-validation model, drawing upon research reports published in peer-reviewed journals since 2000. Through the lens of validity, we consider both the more traditional research themes in the field of SET (i.e., the dimensionality debate, the 'bias' question, and questionnaire design) and some recent trends in SET research, such as online SET and bias investigations into additional teacher personal characteristics. The review provides a clear idea of the state of the art with regard to research on SET, thus allowing researchers to formulate suggestions for future research. It is argued that SET remains a current yet delicate topic in higher education, as well as in education research. Many stakeholders are not convinced of the usefulness and validity of SET for both formative and summative purposes. Research on SET has thus far failed to provide clear answers to several critical questions concerning the validity of SET.

Keywords: student evaluation of teaching, validity, higher education, educational policy

Student evaluation of teaching (SET) is used as a measure of teaching performance in almost every institution of higher education throughout the world (Zabaleta, 2007). Universities and university colleges have developed relatively complex procedures and instruments for collecting, analyzing, and interpreting these data as the dominant or, in some cases, the sole indicator of teaching quality. This widespread use is largely due to the apparent ease of collecting the data and presenting and interpreting the results (Penny, 2003). In addition, students are considered important stakeholders in the process of gathering insight into the quality of teaching in a course, as "the opinions of those who eat the dinner should be considered if we want to know how it tastes" (Seldin, 1993, p. 40). Although SET was originally intended primarily for formative purposes, such evaluations came into use for faculty personnel decisions in the 1970s (Galbraith, Merrill, & Kline, 2012). More recently, SET procedures have been included as a key mechanism in internal quality-assurance processes as a way of demonstrating an institution's performance in accounting and auditing practices (Johnson, 2000).

Purpose of SET

Student evaluation of teaching serves three purposes: (a) improving teaching quality, (b) providing input for appraisal exercises (e.g., tenure/promotion decisions), and (c) providing evidence for institutional accountability (e.g., demonstrating the presence of adequate procedures for ensuring teaching quality; Kember, Leung, & Kwan, 2002). In most institutions, SET is obviously used for formative purposes (e.g., as feedback for the improvement of teaching) as well as for summative purposes (e.g., mapping teaching competence for administrative decision-making and institutional audits; Arthur, 2009; Burden, 2008; Edström, 2008; Emery, Kramer, & Tian, 2003). These dual usages—and the unresolved tension between them (Penny, 2003)—makes the use of SET fragile. On the one hand, teachers are convinced of the value of SET as an instrument for feedback on their teaching (Balam & Shannon, 2010; Griffin, 2001; Kulik, 2001). Results obtained from SET help them to improve the quality of their teaching, as they provide instructors with insight into the strengths and weaknesses of their teaching practice, based on student opinions. For this reason, one can assume that many instructors welcome SET results in order to improve their subsequent teaching. On the other hand, it has been argued that the principal purpose of SET involves its use as a measure for quality monitoring, administrative policymaking (Penny & Coe, 2004), and for determining whether teachers have achieved a required standard in their teaching practice (Bolivar, 2000; Chen & Hoshower, 2003).

This justification for using SET in staff appraisals is related to an increasing focus on internal quality assurance and performance management in universities, which have become subject to the demands of consumer satisfaction (Blackmore, 2009; Olivares, 2003, Titus, 2008). Student satisfaction has come to play an important role in this *managerial approach* (Jauhiainen, Jauhiainen, & Laiho, 2009; Larsen, 2005; Valsan & Sproule, 2005), which is based on such key concepts as accountability, visibility, and transparency (Douglas & Douglas, 2006; Molesworth, Nixon, & Scullion, 2009). Teacher performance and the quality of teaching could thus be defined as the extent to which student expectations are met, thus equating student *opinions* with *knowledge*. For this reason, many faculty members have been questioning the validity and reliability of SET results for many years (Ory, 2001). Their concerns are comprehensible and appropriate as SET results can have serious effects on a teacher's professional career (Kogan, Schoenfeld-Tacher, & Helleyer, 2010).

Teachers' Concerns About the Validity and Reliability of SET

One of the major concerns involves the validity and the reliability of student opinions (i.e., the extent to which students are capable of providing appropriate teacher evaluations). Faculty concerns include the differences between the ways in which students and teachers perceive effective teaching, as well as the relationship of these perceptions to factors that are unrelated to good teaching. In some instances, SET surveys are even known as "happy forms" (Harvey, in Penny, 2003, p. 400) that are used for "personality contests" (Kulik, 2001, p. 10) or as a measure of "customer satisfaction" (Beecham, 2009, p. 135). Second, the sometimes poorly designed questionnaires suggest that the architects of the questionnaire also lack common understanding or consensus regarding what constitutes good or effective teaching (Johnson, 2000; Knapper, 2001). In addition, many instruments are not tested with regard to their psychometric properties (Richardson, 2005).

Third, the common use of SET by means of administering standard questionnaires to be completed (in most cases, anonymously) by all students taking part in a course has been called into question. Administering SET in this way depersonalizes the individual relationship between teachers and their students. For example, Platt (1993) argued that "only the composite opinion of the majority of the students speaks" (p. 5) in a SET report, further warning each student that "you count only as you add to a sum into which you disappear without a trace" (p. 2). Most SET procedures allow little or no space for discussing, explaining, or negotiating the results with the students (Johnson, 2000). Fourth, the interpretation of SET results is more complicated than it looks, and it entails a risk of inappropriate use by both teachers and administrators for both formative and summative purposes (Franklin, 2001). Fifth, many faculty members are unaware of the sheer volume of research on SET (in which almost all of their concerns are addressed; Ory, 2001). It has been shown, however, that teachers who are familiar with the SET literature are more positive toward such evaluations (Franklin & Theall, in Paulsen, 2002). This lack of familiarity with the literature has generated a number of persistent myths or urban legends concerning SET, most of which have been invalidated in many research reports (Aleamoni, 1999).

Given these concerns, it is not that surprising that many teachers fear their next SET reports, even though they tend to see SET as useful for summative decision-making (Beran & Rokosh, 2009). In some cases, this leads to practices aimed at increasing SET scores rather than improving instruction (Simpson & Siguaw, 2000). The tyranny of the evaluation form may lead to grading leniency, which can result in grade inflation (Crumbley, Flinn, & Reichelt, 2010; Eiszler, 2002; Ellis, Burke, Lomire, & McCormack, 2003; Langbein, 2008; Oleinik, 2009; Redding, 1998). At the same time, many valuable thoughts and suggestions from students remain untouched, as faculty members who do not perceive SET instruments as valid measurements tend to ignore the results (Simpson & Siguaw, 2000).

Research on SET

As mentioned above, most stakeholders (e.g., teachers, students, administrators, and policymakers) are unaware of the number of research studies that have been conducted within the domain of SET. Several thousands of research studies have appeared since the publication of the first report on SET by Remmers and Brandenburg in 1927, addressing various elements of these evaluations. Nevertheless, the primary focus of these studies is on the validity of student opinions and their relationship to possible biasing factors (for overviews, see Aleamoni, 1999; Marsh, 1984, 1987, 2007b; Marsh & Roche, 1997; Wachtel, 1998). Although the majority of research shows that SET provides useful information to both teachers and administrators (Marsh, 1987; Ory, 2001; Penny, 2003), the validity of such evaluations continues to be called into question (Clayson, 2009).

Several authors (Olivares, 2003; Ory & Ryan, 2001; Onwuegbuzie, Daniel, & Collins, 2009) have developed conceptual validity frameworks for assessing the validity of SET (e.g., regarding the extent to which scores generated by SET instruments measure the variables they are intended to measure). These frameworks are based on Messick's (1989, 1995) unified conceptualization of validity. Onwuegbuzie et al. (2009) developed a meta-validity model, which is subdivided to address construct, content, and criterion validity. Each of these types of validity is subdivided into areas of evidence. *Construct-related validity* (substantive validity, structural validity, convergent validity, discriminant validity, divergent validity, outcome validity, generalizability) addresses the extent to which an instrument can be seen as a meaningful measure of a given characteristic. *Content-related validity* (face validity, item validity, sampling validity) concerns the extent to which the items of an instrument are appropriate representations of the content being measured. *Criterion-related validity* (concurrent validity, predictive validity) is associated with the extent to which scores are related to another independent and external variable that can serve as a direct measure of the underlying characteristic.

The Current Study

The purpose of this article is to provide a systematic overview of the recent literature on SET (since 2000) using the meta-validity model for assessing the score validity of SET designed by Onwuegbuzie et al. (2009). Through this validity lens, we consider both the more traditional research themes in the field of SET (i.e., the dimensionality debate, the "bias" question, and questionnaire design) and some recent trends in SET research such as online SET and bias investigations into additional teacher personal characteristics. Our goal is to summarize the state of the art in SET research and provide a basis for developing ideas for future research.

Method

Literature Search

Given the inconsistent use of terminology concerning SET, the literature search for this study was based on a variety of terms that refer to the concept of SET (i.e., questionnaire-based student evaluations of an individual course). The following keywords were used (separately and in combination) when searching the electronic databases, Web of Science, EBSCO, and ERIC: *SET, student evaluation of teaching, student ratings, student ratings of instruction, teacher evaluation, teaching effectiveness, teaching performance, higher education*, and *student evaluations*. To ensure that the search would generate an overview of the state of the art in high-quality research concerning SET, the search was limited to articles published in international peer-reviewed journals since 2000. In the supporting texts, however, we will also discuss some classic studies published prior to 2000, which cannot be ignored.

We read the abstracts of 542 peer-reviewed journal articles. Each abstract was read by at least two authors to determine the article's relevance to the review (based on its relationship with validity issues regarding SET, methodology, and conclusions). The search was not limited to empirical studies but also included conceptual, theoretical, and review studies since such papers draw important conclusions for SET and SET research as well. The database search left us with 210 articles that were fully read by the first author. The snowball method was then used to identify additional works (including chapters in edited books) through the references listed in the selected articles.

For each article, specific information was noted, including (a) authors, (b) year of publication, (c) journal, (d) objectives of the study, (e) methodology, (f) important findings and conclusions, and (g) relevance for this review (i.e., the validity of SET). In case of disagreements concerning important issues such as methodology, findings, and relevance to the review, an article was read by the other authors and discussed at a meeting. Based on the discussion, a decision was made to include or exclude that article. Although the literature search was limited to articles published in the English language, this review has an international character, as it includes 31 articles written by authors residing in 11 countries other than the United States.

The final database consisted of 160 pieces (158 journal articles and 2 book chapters), including empirical studies, theoretical pieces, and other types of articles. An initial reading of all articles suggested that each of the selected studies could be classified as addressing at least one of the aforementioned types of validity. The following sections provide a narrative review of the recent SET literature, organized according to the meta-validation model designed by Onwuegbuzie et al. (2009). In the reference list, all studies included in the review are indicated with an asterisk.

Results

Content-Related Validity

Sampling validity and item validity. Although SET has become common practice in many institutions, and although it has been the subject of thousands of research studies, there is a surprising amount of variation in the SET instruments used to collect feedback from students. The starting point seems simple: Institutions need instruments that will allow them to gather information (preferably comparable) for different types of courses as quickly as possible. Such surveys must also be highly economical (Braun & Leidner, 2009). Although Lattuca and Domagal-Goldman (2007) advocated the use of qualitative methods in SET, in practice, such evaluations usually consist of standardized questionnaires (including both rating scales and open-ended items) aimed at providing a descriptive summary of the responses for both the teacher and the teacher's department head, as well as the institution's educational board or personnel system (Richardson, 2005). Nevertheless, this dual objective has generated a panoply of SET instruments that vary greatly in both content and construction, due to the characteristics and desires of particular institutions. This variety has implications for the item validity (i.e., the extent to which SET items are decent representations of the content area) and the sampling validity (i.e., the extent to which the SET instrument as a whole represents the whole content area) of SET instruments.

Several well-designed and validated instruments are available, however, including the Instructional Development and Effectiveness Assessment (IDEA; Cashin & Perrin, 1978), the Students' Evaluation of Education Quality (SEEQ; Marsh, 1982; Marsh et al., 2009), the Course Experience Questionnaire (CEQ; Ramsden, 1991), the Student Instructional Report (SIR II; Centra, 1998), and the Student Perceptions of Teaching Effectiveness (SPTE; Burdsal & Bardo, 1986; Jackson et al., 1999), as well as the more recent Students' Evaluation of Teaching Effectiveness Rating Scale (SETERS; Toland & De Ayala, 2005), the Student Course Experience Questionnaire (SCEQ; Ginns, Prosser, & Barrie, 2007), the Teaching Proficiency Item Pool (Barnes et al., 2008), the SET37 questionnaire for student evaluation of teaching (SET 37, Mortelmans & Spooren, 2009), the Exemplary Teacher Course Questionnaire (ECTQ; Kember & Leung, 2008), and the Teaching Behavior Checklist (Keeley, Furr, & Buskist, 2010; Keeley, Smith, & Buskist, 2006). Validation procedures for other instruments have not been successful (Haladyna & Amrein-Beardsley, 2009).

Still, many instruments are developed without any clear theory of effective teaching (Ory & Ryan, 2001; Penny, 2003). They therefore lack any evidence of content validity and thus might fail to measure what they claim to measure (Onwuegbuzie et al., 2009). A clear understanding of effective teaching is a prerequisite for the construction of SET instruments. Although it is logical to assume that educational scientists have reached some level of consensus regarding the characteristics of *effective teachers* (e.g., subject knowledge, course organization, helpfulness, enthusiasm, feedback, interaction with students), existing SET instruments vary widely in the dimensions that they capture. In a theoretical article on the shortcomings of SET research, Penny (2003) argued in favor of establishing an interinstitutional task force to formulate a list of standards or characteristics within a common framework of effective teaching, which can be used as a basis for the development of SET instruments. We add two conditions: (a) institutions should be able to select the aspects that are most important, according to their educational vision and policy, thereby developing SET instruments that are consistent with their own preferences; and (b) all stakeholders (i.e., administrators, teachers, and students) should be involved in the definition of these characteristics.

Face validity. The latter condition is derived from the growing body of research showing that SET instruments, which are usually designed by administrators (based on some didactic model of teaching), do not always reflect the students' perspective concerning effective teaching. This disconnect affects the face validity of SET instruments (i.e., the extent to which the items of a SET instrument appear relevant to a respondent). For this reason, the results of such evaluations might be biased, as students tend to respond to items according to their own conceptions of good teaching (Kember, Jenkins, & Kwok, 2004). Kember and Wong (2000), for instance, concluded from interviews with 55 Hong Kong undergraduate students that students' perceptions of teaching quality should be seen as the result of an interplay between students' conceptions of learning (a continuum between active and passive learning) and students' beliefs about teaching of the lecturer (ranging between transmissive and nontraditional teaching). Besides, based on a sequential mixed-method analysis that led to a model that represented four meta-themes and nine themes that (according to 912 students) reflected students' conceptions of effective college teaching, Onwuegbuzie et al. (2007) concluded that three of these themes were not represented in the teaching-evaluation forms used at their university (student centered, expert, and enthusiast).

Bosshardt and Watts (2001) showed that, although the perceptions of students and teachers with regard to effective teaching are positively correlated, differences exist as well. For example, students care more about the teacher's preparation for class than instructors do. Pan et al. (2009) analyzed both quantitative (student ratings) and qualitative (students' comments in open-ended questions) student feedback data and found that, contrary to popular perception, students value the quality of teaching (e.g., ability to explain, aiding understanding) more than they value particular instructor characteristics (e.g., humor, a charismatic personality, or storytelling skills). Barth (2008) concurred, having found that students' overall instructor ratings are driven primarily by the quality of instruction. Factor analysis and multiple regression analysis (167 classes, 30 instructors, +4,000 students) revealed that each of five factors (quality of instruction, course rigor, level of interest, grades, and instructor helpfulness) had a strong statistically significant relation with the overall instructor rating (with the five factors explaining 95% of the variance in the measure of overall instructor rating). Using multigroup SEM on a sample of 3,305 first-year and third-year undergraduate students in Hong Kong, Kember and Leung (2011) showed that students from four different disciplines (humanities, business, hard science, health sciences) shared the same ideas concerning the nature of an effective teaching and learning environment. There were nevertheless differences among disciplines concerning the extent to which some elements within this environment were brought into play. Pozo-Munoz, Rebolloso-Pacheco, and Fernandez-Ramirez (2000) used factor analysis based on data from a 39-item semantic differential scale to define the attributes of the *ideal teacher*, according to 2,221 students from a Spanish university. The most valued teacher characteristics were having knowledge, having adequate communication skills, and being competent in teaching.

Goldstein and Benassi (2006) noted that SET scores are higher when students and teachers agree on the characteristics of excellent lecturers. Based on a study that involved both students' and their teachers' conceptions of the *ideal teacher* and students' perceptions of teaching quality, they found that mean SET scores were higher (6.00 on a 7-point scale) in the no-discrepancy group (i.e., where students' and teachers' conceptions of the ideal teacher coincided) compared to the positive (when students rated the items on the ideal lecturer scale as more important than did their teacher) and negative (when teachers rated the items on the ideal lecturer scale as more important than did their students) discrepancy groups (mean SET scores were 5.52 and 5.68, respectively). ANOVA results showed a reliable quadratic effect (Cohen's $d = .26$) between the SET scores from these three groups. Kember and Leung (2008) derived nine principles of *good teaching* from interviews with award-winning teachers about their insights and practices. These principles form the basis for their Exemplary Teacher Course Questionnaire.

In summary, the research literature suggests that there is a risk that important SET stakeholders (i.e., teachers, students, and questionnaire architects) may differ in their conceptions with respect to effective teaching and, thus, should be involved in the process of defining good teaching, as well as in the design of SET instruments.

Construct-Related Validity

Structural validity and the dimensionality debate. Although it is widely accepted that SET should be considered multidimensional (given that teaching consists of many aspects) and that SET instruments should capture this multidimensionality, many authors and institutional boards argue in favor of single, global scores (Apodaca & Grad, 2005). Important questions thus arise with regard to the following: (a) the number and dimensions of effective teaching that can be distinguished and (b) the possibility of compiling an overall score based on these dimensions.

The SET literature reflects no consensus on the number and the nature of dimensions (Jackson et al., 1999). This lack of consensus is due to conceptual and methodological problems, given that (a) we lack a theoretical framework concerning effective teaching, (b) views on effective teaching differ both across and within institutions (Ghedin & Aquario, 2008), and (c) the measurement of dimensions continues to be relatively data-driven (with different post hoc analyzing techniques and different decision rules), with a few exceptions. The latter observation calls into question the structural validity of SET instruments (i.e., the extent to which the factors measured by a SET-instrument are consistent with the factor structure of the construct). Onwuegbuzie et al. (2009) argued that this method of assessing the dimensions of instruction does not guarantee that items included in SET forms represent effective teaching; instead, they should be seen as indicators of teaching performance (as perceived by the students).

Table 1 provides an overview of the dimensions captured in recently reported SET instruments, thereby demonstrating the wide variety that exists with regard to the aspects of teaching and course quality that are measured in SET. Feedback from students regarding particular aspects of courses is helpful as a guide for improving teaching. Teachers receive precise and detailed suggestions for refining their teaching in a particular course. Because SET is used for administrative decision-making as well, however, there is a need for a unidimensional and global SET score that provides a clear measure of overall teaching quality (McKone, 1999). In the 1990s, several leading SET authors entered into debate with regard to the dimensionality of SET. This debate also addressed the important question of whether SET scores on several dimensions could be captured by a single-order factor that represents a global construct (i.e., “general instructional skill”) and whether such a factor could be used for summative purposes (see, e.g., Abrami & d’Apollonia, 1990, 1991; Marsh, 1991b; Marsh & Hovecar, 1991). The debate resulted in a compromise, which recommends the use of both specific dimensions and global measures for administrative decision-making, using the weighted averages of individual dimensions to generate an overall rating (Marsh, 1991a). Recent research provides further evidence on this matter. Many authors report evidence to support the multidimensionality of teaching, furnishing proof of higher order factors that reflect general teaching competency (Apodaca & Grad, 2005; Burdsal & Harrison, 2008; Cheung, 2000; Harrison, Douglas, & Burdsal, 2004; Mortelmans & Spooren, 2009).

Relationships between several dimensions of SET have been studied as well, using structural equation modeling. For example, Paswan and Young (2002) reported that the factors Course Organization and Student-Instructor Interaction have a positive effect on the factors Instructor Involvement (.66 and .78, respectively) and Student Interest (.60 and .65), on the 21-item Student Instructional Rating System (SIRS) instrument, whereas the factor Course Demands has a negative effect on these factors (–.38 and –.43). The authors argued that relationships between the factors in a SET instrument should be considered when interpreting the results. In a similar study, Marks (2000) reported that some constructs have large effects on others. For instance, students’ ratings of teaching ability were affected by their expectations regarding the fairness of grading (.24). Marks concluded that SET may lack discriminant validity (see below) and advised caution when using global SET measures for summative decisions. Gursoy and Umbreit (2005) provided evidence for a model in which students’ perceptions regarding the

TABLE 1
Summary of dimension numbers in SET instruments (ever since 2000)

Author	Instrument	N° of Dimensions	Dimensions
Barth (2008)	Institutional	5	Quality of instruction Course rigor Level of interest Grades Instructor helpfulness
Cohen (2005)	Institutional	2	Course Teacher
Ginns et al. (2007)	SCEQ	5	Good teaching Clear goals and standards Appropriate assessment Appropriate workload Generic skills
Gursoy & Umbreit (2005)	Institutional	4	Organization Workload Instruction Learning
Keeley et al. (2006) Keeley et al. (2010)	TBC	2	Caring and supportive Professional competency and Communicational skills
Kember & Leung (2008)	ETCQ	9	Understanding fundamental content Relevance Challenging beliefs Active learning Teacher–student relationships Motivation Organization Flexibility Assessment
Marks (2000)	Initial instrument	5	Organization Workload/difficulty Expected/fairness of grading Instructor liking/concern Perceived learning
Marsh et al. (2009) Marsh (1982) Coffey & Gibbs (2001)	SEEQ	9	Learning/value Instructor enthousiasm Organization/clarity Group interaction Individual rapport Breadth Exam/graded materials Readings/assignments Workload difficulty

(continued)

TABLE 1 (continued)

Author	Instrument	N° of Dimensions	Dimensions
Mortelmans & Spooren (2009) Spooren (2010)	SET37	12	Clarity of objectives Value of subject matter Build-up of subject matter Presentation skills Harmony organization course-learning Course materials Course difficulty Help of the teacher during the learning process Authenticity of the examination(s) Linking-up with foreknowledge Content validity of the examination(s) Formative evaluation(s)
Shevlin, Banyard, Davies, & Griffiths (2000)	Initial instrument	2	Lecturer ability Module attributes
Toland & De Ayala (2005)	SETERS	3	Instructor's Delivery of Course Information Teacher's Role in Facilitating Instructor/Student Interactions Instructor's Role in Regulating Students' Learning

Note. Keeley et al. (2006) found a good fit for one-factor model to the data as well. ETCQ = Exemplary Teacher Course Questionnaire; SCEQ = Student Course Experience Questionnaire; SEEQ = Students' Evaluation of Education Quality; SETERS = Students' Evaluation of Teaching Effectiveness Rating Scale; TBC = Teaching Behavior Checklist..

organization, course workload, and instructional abilities of their teachers have a positive impact on a fourth construct, their perception of learning (the estimated standardized path coefficients were .32, .04, and .60, respectively, $R^2 = .78$).

In summary, SET researchers agree that SET and SET instruments should capture multiple aspects (dimensions) of good teaching practice. Due to the absence of an agreement with respect to the number and the nature of these dimensions, which should be based on both the theory and empirical testing, SET instruments vary greatly in both the content and the number of dimensions. Additionally, recent research has revealed that many dimensions in SET instruments seem to be affected by a global (unidimensional) construct, which could be used for summative purposes. Thus, on the one hand, one could use the results on one or more particular dimensions when working on the improvement of (teaching) a course. On the other hand, an overall score derived from the (weighted) scores on dimensions of which it is known that they belong can be used to create a general factor representing general teaching competency, which in turn can be used for the evaluation of teaching staff.

Convergent validity. The most common method for assessing the convergent validity of SET instruments is to examine the relationship of SET scores to student achievement (objective measure) or student perceptions of learning (subjective measure), which are considered proxies for the students' actual learning. Reviews and multisection studies suggest positive and moderate correlations between *student grades* and SET scores (Onwuegbuzie et al., 2009), varying between .10 and .47 (Cohen, 1981; Feldman, 1997). These studies also provide evidence regarding the criterion-related validity (concurrent validity) of SET.

Recent studies by Braun and Leidner (2009) and by Stapleton and Murkison (2001) indicate moderate to strong statistically significant associations between *students' self-reported acquisition of competence* and their satisfaction with teaching behavior. In these studies, correlation coefficients ranged between .28 and .75. Based on a meta-analysis of the literature (with a majority of the studies conducted in the 1970s), Clayson (2009) found a small average relationship (.13) between students' learning (i.e., *testing results*) and SET. Galbraith et al. (2012) suggested that the relationship between student achievement (as measured by a standardized learning-outcome test) and SET scores is nonlinear, with the most effective teachers falling within the middle percentiles of SET scores. Other researchers have found little or no support for the validity of SET as a predictor of student learning (e.g., Mohanty, Gretes, Flowers, Algozzine, & Spooner, 2005; Stark-Wroblewski, Ahlering, & Brill, 2007).

In this regard, however, it is appropriate to question the ways in which student achievement has been measured in previous studies. Student perceptions of learning might not always reflect actual learning (e.g., students could think that they had learned a lot during a course, even if they failed the examinations). And because student outcomes on objective tests are affected by other factors as well (e.g., prior knowledge, interest in the subject matter), they cannot be considered precise measures of actual student learning in a course. For this reason, a pretest is needed at the beginning of the course to estimate accurately how much learning individual students acquired at the end of the course. Students who are already familiar with the subject matter might receive good grades even though they do not learn very much, whereas slower students might fail the examinations even though they achieve considerable learning progress during the course. Future research using pretests and posttests of student achievement can provide useful insights into discussions of the relationship between student learning and SET.

Most authors agree that SET is correlated with teachers' self-evaluations, alumni ratings, and evaluations by trained observers (Marsh, 1987; Richardson, 2005; Roche & Marsh, 2000). This finding provides further evidence supporting the convergent validity of SET. Renaud and Murray (2005) reported a moderately strong correlation (.54) between SET and actual teaching behavior, as observed from videotapes. Given the relatively small correlations between SET and peer or administrator ratings, it is important to consider that SET is only one of many instruments available for mapping teaching effectiveness (Marsh & Roche, 1997). On many campuses, however, SET is used as an important (and, in some cases, the sole) indicator of teaching quality in personnel decisions, implying that only one important stakeholder is involved in the evaluation process. Given the risk of differences among stakeholders regarding the concept of teaching effectiveness, and given that the persistent feelings of teachers that student evaluations may be biased by external characteristics, we argue that personnel files should include other measures of teaching quality (e.g., teachers' reflection on their SET scores, observation reports by peers or educational experts) as well.

Several authors (Burdsal & Harrison, 2008; Emery et al., 2003) also argue in favor of teaching portfolios, which contain various indicators of teaching performance, with student evaluations as one component. At the institutional level, SET can be included as one indicator (e.g., in addition to student progress and retention rates) when using DEA (Data Envelopment Analysis) to explore an institution's educational performance using the learning performance of its students (Montoneri, Lee, Lin, & Huang, 2011; Montoneri, Lin, Lee, & Huang, 2012).

In summary, the research literature revealed the existence of (small to strong) positive correlations between SET scores and student achievement, expert ratings of teaching behavior, self-ratings, and alumni ratings. These results provide evidence of the convergent validity of SET. However, due to the variety in stakeholders' views concerning good teaching and due to the variety in the measurement of student achievement, SET should not be the only indicator of teaching effectiveness in personnel files.

Discriminant validity and divergent validity. Many recent SET studies continue to address the question of bias, or the effect of factors that are not necessarily related to teaching quality on SET scores (Centra & Gaubatz, 2000). This issue involves the discriminant validity and divergent validity of SET, which has received considerable attention from researchers, administrators, and teachers. Although most leading SET researchers are convinced of the validity of SET, as research has found potentially biasing factors to be of little or no influence (Centra, 2003; Marsh & Roche, 2000), bias studies continue to play a central role in the recent literature.

Table 2 provides an overview of recent studies that address student-related, teacher-related, and course-related characteristics that might affect SET. Although it is not our intention to discuss each of these studies, it is clear that not all of the reported characteristics should be considered biasing factors. Some are meaningful indicators of student learning and are therefore logically related to effective teaching and SET. For example, student effort and class attendance indicate the interest and motivation of students in a particular course and are at least partly dependent upon the organization of and the teaching in that course. The experience, rank, and research productivity of the teacher are valuable indicators of a teacher's educational skills and knowledge of the subject matter.

On the other hand, although the course discipline and the sexual orientation of the teacher have nothing to do with effective teaching, they could be biasing factors for SET. The same applies to the teacher's gender or race. Further discussion concerns whether several other variables should be interpreted as biasing factors. For example, the relationship of SET to both course workload and student grade expectations continue to provoke discussions among SET researchers (for

TABLE 2

Relationships between student, teacher, and course characteristics and SET scores

Characteristic	Author(s)	Measure	Significant?	Interpretation
Student				
Student's cognitive background	Ting (2000)	Student's major and year of enrollment	Y	Mature students majoring in the same subject as the course, give higher SET
Class attendance	Beran & Violato (2005)	Frequency of attendance in the course	Y	Students who attend most classes (because of interest, motivation, being likely to learn, etc.) provide higher SET
	Davidovitch & Soen (2006a)		Y	
	Spooren (2010)		Y	
Students' effort	Heckert, Latier, Ringwald-Burton, & Drazen (2006)	Student effort (i.e., preparation for class, in-class behavior, etc.)	Y	Teachers who encourage students to make more effort, get higher SET
Expected grade	Beran & Violato (2005)	Student's expected grade	Y	The higher the expected grade, the higher SET
	Griffin (2004)		Y	
	Guinn & Vincent (2006)		Y	
	Langbein (2008)		Y	
	Maurer (2006)		Y	
	McPherson (2006)		Y	
	McPherson & Todd Jewell (2007)		Y	
	McPherson, Todd Jewell, & Kim (2009)		Y	
	Olivares (2001)		Y	
	Remedios & Lieberman (2008)		Y	

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
	Stapleton & Murkison (2001)		Y	
	Marsh & Roche (2000)	Student’s expected grade	Y	The higher the expected grade, the higher SET. But some SET factors are unrelated to expected grade, and relationship grade–SET is nonlinear (the highest grades are not correlated with SET)
	Isely & Singh (2005)	Expected grade at the class level	Y	SET are higher in classes in which students expect higher grades
	Centra (2003)	Student’s expected grade	N	
	Stodnick & Rogers (2008)		N	
Final grades	Langbein (2008)	Student’s final grade	Y	The higher the grade, the higher SET
	Spooren (2010)		Y	
Study success	Spooren (2010)	Passing the examinations in one or two times	Y	Students who had to retake the examinations for the course, give lower SET
Student’s gender	Basow, Phelan, & Capotosto (2006) Centra & Gaubatz (2000)	Student’s gender and teacher’s gender	Y	There seem to be some gender preferences (i.e., female students give higher ratings to female teachers)
	Kohn & Hartfield (2006)	Student’s gender and teacher’s gender	Y	Female students give higher SET than male students Female students give higher SET to male teachers than male students
	Santhanam & Hicks (2001)	Student’s gender	Y	Female students give higher SET than male students
	Smith, Yoo, Farr, Salmon, & Miller (2007)		Y	

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
	Spooren (2010)		N	
Student's goals	Remedios & Lieberman (2008)	Student's goal orientation (i.e., competitive, mastery, etc.)	Y	Students with a mastery goal are more likely to give positive SET
Student's age	Spooren (2010)	Students' age	Y	The greater the age, the higher SET
Grade discrepancy	Griffin (2004)	Difference between expected grade and believed deserved grade	Y	Students tend to punish teachers when expected grades are lower than they believed to deserve
Grading leniency	Griffin (2004)	Student's perception of instructor's grading	Y	The more lenient the grading, the higher SET
	Olivares (2001)		Y	
Pre-course interest	Olivares (2001)	Level of interest in the course	N	
Interest change during the course	Olivares (2001)	Interest change (increased, decreased, stable)	Y	Interest change during the course is positively associated with SET (increased interest leads to higher SET)
Precourse motivation	Griffin (2004)	Desire to take the course	Y	The stronger the desire to take the course, the higher SET
Teacher				
Instructor's gender	Basow & Montgomery (2005)	Teacher's gender	Y	Female teachers receive higher SET
			Y	Female teachers receive higher SET
	Smith et al. (2007)		Y	Male teachers receive higher SET
	McPherson et al. (2009)		N	
	McPherson & Todd Jewell (2007)			
Instructor's reputation	Griffin (2001)	Instructor reputation as perceived by the students	Y	Teachers with a positive reputation receive higher SET

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
Research productivity	Stack (2003)	Citations and post-PHD year	Y	The better a teacher's quality of research, the higher SET
	Ting (2000)	Number of publications	N	
Instructor's teaching experience	McPherson et al. (2009)	Total semesters of teaching experience	Y	More experienced teachers receive higher SET
	McPherson & Todd Jewell (2007)	Teaching experience (<5, 5–10, 11+ semesters)	Y	
	McPherson (2006)		Y	
Instructor's age	McPherson et al. (2009)	Teacher's age	Y	Younger teachers receive higher SET
	Spooren (2010)		N	
Instructor's language background	Ogier (2005)	English as a second language (ELS) vs. native speakers	Y	ELS speakers receive lower SET than native speakers (especially in the science faculties)
Instructor's race	McPherson et al. (2009)	Teacher's race	Y	White teachers receive higher SET in upper-level courses
	McPherson & Todd Jewell (2007)		Y	
Instructor's tenure	McPherson & Todd Jewell (2007)	Tenured vs. nontenured faculty	Y	Nontenured faculty receive lower SET
Instructor's rank	McPherson et al. (2009)	Adjunct instructors vs. tenure-track faculty	Y	Adjunct instructors receive higher SET than tenure-track faculty
	Spooren (2010)		Y	
	Ting (2000)	Full professors vs. professors, associate professors, lecturers, and junior lecturers Senior lecturers vs. all other ranks	N	(Full) professors receive higher SET than associate professors and professors

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
Instructor's sexual orientation	Ewing, Stukas, & Sheehan (2003)	Sexual orientation (gay/lesbian vs. unspecified)	Y	After *strong* lectures, known gay/male teachers receive lower SET, but after *weak* lectures they receive higher SET
Instructor's personal traits	Shevlin et al. (2000)	Teacher charisma	Y	A modeled "charisma" factor explains 69% and 37% of the variation in the "lecturer ability" and "module attributes" factors, respectively
	Clayson & Sheffet (2006)	Teacher personality (Big Five)	Y	Students' evaluations of their instructor's personality (Big Five) show significant correlations with SET
	Patrick (2011)		Y	
	Campbell, Gerdes, & Steiner (2005)	Physical attractiveness	N	
	Feeley (2002)		Y	Measures of instructor physical attractiveness have significant relationships with measures of effective teaching
	Gurung & Vespia (2007)		Y	Likable, good-looking, well-dressed, and approachable teachers receive higher SET
	Hamermesch & Parker (2005)		Y	Good-looking teachers receive higher SET (besides, the impact is larger for male than for female instructors)
	Riniolo, Johnson, Sherman, & Misso (2006)		Y	Professors perceived as attractive received student evaluations about 0.8 of a point higher on a 5-point scale
	Wendorf & Alexander (2005)	Instructor fairness	Y	SET is significantly related to perceptions of the fairness of grading procedures, the fairness of instructor–student interactions, and the fairness of the expected grades

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
	Kim, Damewood, & Hodge (2000)	Professor attitude	Y	Instructors who are perceived as approachable, respectful, pleasant … receive higher SET
	Dunegan & Hrivnak (2003)	Image compatibility	Y	SET scores are significantly related to image compatibility (i.e., the comparison between an image of an "ideal" instructor with an image of the instructor in this course)
	Delucchi (2000)	Instructor likability	Y	Instructors who are rated high in likability receive higher SET
	Tom, Tong, & Hesse (2010)	Initial impressions of a teacher	Y	SET based upon 30-s video clips of instructors in the classroom correlate strongly with end of the term SET
Course				
Class size	Bedard & Kuhn (2008)	Class size	Y	Nonlinear, negative relationship between class size and SET (relationship becomes stronger for higher class sizes)
	McPherson (2006)		Y	
	McPherson et al. (2009)		N	
	Ting (2000)		N	Negative relationship between class size and SET
Class attendance rate	Ting (2000)	Class attendance rate (ratio of students present in evaluation exercise and the class size)	Y	The higher the class attendance rate, the higher SET
Class heterogeneity	Ting (2000)	Index of diversity (based on students' years of enrolment in the same class)	N	

(continued)

TABLE 2 (continued)

Characteristic	Author(s)	Measure	Significant?	Interpretation
Course difficulty	Remedios & Lieberman (2008)	Student's perceived course difficulty	Y	The more difficult the course, the lower SET
	Ting (2000)	Identified by institution	N	
Course discipline	Basow & Montgomery (2005)	Course discipline	Y	Natural science courses receive lower SET
	Beran & Violato (2005)		Y	Natural science courses receive lower SET
Course workload	Centra (2003)	Student's perception of course workload	Y	SET are lower for both difficult and too elementary courses; "just right" courses receive the highest SET
	Marsh & Roche (2000)		Y	A positive relationship between course workload and SET
	Marsh (2001)		Y	A positive, nonlinear relationship between good (useful) workload and SET (relationship becomes smaller for higher workloads)
	Dee (2007)		N	
Course level	Santhanam & Hicks (2001)	Course's year level	Y	SET in higher year level are more positive
Course type	Beran & Violato (2005)	Lab-type vs. lectures/tutorials	Y	Lab-type courses receive higher SET
Elective vs. required courses	Ting (2000)	Required vs. elective courses	Y	Elective courses receive higher SET (lecturing performance)
General education vs. specific education	Ting (2000)	General vs. specific course contents	Y	Courses with specific content matters receive higher SET
Syllabus tone	Harnish & Bridges (2011)	Friendly vs. unfriendly syllabus tone	Y	Teachers with a friendly written syllabus tone receive higher SET

overviews, see Brockx, Spooren, & Mortelmans, 2011; Griffin, 2004; Gump, 2007; Marsh, 2001, 2007b). Many SET studies provide evidence to support the validity hypothesis with regard to interpreting the relationship between expected grades and SET, thus suggesting that the positive relationship between expected grades and SET has to do with the fact that students who have learned a great deal—and who thus expect good grades—assign higher SET scores for their teachers. Such studies have also rejected the hypothesis concerning the existence of a negative (and thus biasing) relationship between course workload and SET (Marsh, 2001; Marsh & Roche, 2000). Nevertheless, other authors continue to advocate the grading-leniency hypothesis (i.e., teachers can *buy* good evaluations by giving high grades; see, e.g., Isely & Singh, 2005; Langbein, 2008; McPherson, 2006; McPherson & Todd Jewell, 2007), drawing upon attribution theories and measures of the instructor's grading leniency (as perceived by students) to support their argument (Griffin, 2004; Olivares, 2001).

In addition to research on the impact of the classic and potentially biasing factors, a considerable amount of research focuses on the impact of psychological dynamics on SET. First, some authors argue for the possibility of halo effects in SET. A halo effect can be understood as "a rater's failure to discriminate among conceptually distinct and potentially independent aspects of a ratee's behaviour" (Feeley, 2002, p. 226). The contention is that students base their evaluations of a given teacher or course on a single characteristic of that teacher or course, subsequently generalizing their feelings about this characteristic to most or all other unrelated characteristics of the teacher or course. Shevlin et al. (2000) defined a charisma factor that explains a large portion of the variance in several factors (69% and .39% in the factors Lecturer Ability and Module Attributes, respectively) included in their SET instrument. Significant correlations (ranging between .28 and .72) have been observed among all measures in the SET instrument developed by Feeley (2002), which also includes irrelevant measures (e.g., physical attractiveness).

Ever since Ambady and Rosenthal (1993) found that students' opinions about teachers are formed within seconds of being exposed to the nonverbal behavior and physical attractiveness of these teachers, bias studies have also focused on other personal traits that are considered strongly related to SET. Examples include teacher personality as measured by the Big Five personality traits (Clayson & Sheffet, 2006; Patrick, 2011), physical attractiveness (Campbell et al., 2005; Gurung & Vespia, 2007; Hamermesch & Parker, 2005; Riniolo et al., 2006), instructor fairness (Wendorf & Alexander, 2005), professor attitude (Kim et al., 2000), image compatibility (Dunegan & Hrivnak, 2003), instructor likability (Delucchi, 2000), and initial impressions of a teacher (Tom et al., 2010).

Generalizability. Most of the contradictory research results on SET are due to the great variety of methods, measures, controlling variables, SET instruments, and populations used in these studies. This high degree of variation calls the generalizability of these results into question and makes it almost impossible to make statements concerning, for example, the global effect size of the concurrent validity coefficients with student achievement, or the strength of the relationship of various possibly biasing effects on SET scores. However, several researchers have found that the effect of the possibly biasing factors on SET is relatively small. For instance, Beran and Violato (2005) found that various students and characteristics explained only 7% of the total variance in SET scores. Spooren (2010) reported small local effect sizes of 6.3% for students' grades and of 1.6% for the examination wherein the course grade was given (students that had to retake examinations give lower SET) on SET. The PRV (proportional reduction in variance statistic) for other student, course, and teacher characteristics was estimated close to 0. Smith et al. (2007) noted statistically significant effects of sex of students and sex of instructors on SET scores, but these predictors did not account for more than 1% of the explained variance in SET. These findings suggest that SET outcomes depend primarily upon teaching behavior (Barth, 2008; Greimel-Fuhrmann & Geyer, 2003).

Nevertheless, some authors recommend adjusting raw SET scores in order to purge them of any known biasing effect, especially when these results are used for ranking (McPherson, 2006; McPherson et al., 2009; Santhanam & Hicks, 2001). In this regard, future SET research could also explore the simultaneous administration of SET and such measures as the Marlowe-Crowne Social Desirability Bias Index (Crowne & Marlowe, 1960). This strategy might improve the adequacy of SET for making evaluative decisions, as it would allow the elimination of one type of bias from analyses.

Substantive validity. One crucial topic in the debate on the construct-related validity of SET concerns student behavior when completing SET questionnaires. This issue affects the substantive validity of SET instruments (i.e., the extent to which an instrument is consistent with the knowledge, skills, and processes that underlie a respondent's scores). Understanding how students react to certain questions (or types of questions) and being aware of response patterns provide information that could be useful in the construction of SET items and could increase the substantive validity of SET scores. Recent research has paid considerable attention to what should and should not be done when developing SET questionnaires that take into account the knowledge and skills supposed to underlie students' SET scores.

Instruments used in SET measure students' attitudes toward effective teaching, which should be seen as a latent construct. Such a construct is not immediately observable using a single-item approach that, although sometimes resulting in highly stable estimates (Ginns & Barrie, 2004), assumes that all aspects or dimensions of teaching quality can be observed unequivocally. Spooren, Mortelmans, and Denekens (2007) argued in favor of using Likert-type scales in which sets of items measure several dimensions of teaching quality. These scales allow a straightforward quality check (e.g., by calculating alpha statistics) for each dimension contained in a SET report. Multiple-item scales also provide both the administrator and the teacher with information on score reliability for each particular course evaluation.

Most SET instruments use Likert-type scales to gather information on the quality of teaching in particular courses. This choice is related to ease of use (for both administrators and teachers), given that scales grant a quick and clear view of student opinions regarding the teaching in a particular course. As many authors have observed, however, SET results are subject to bias due to both the content and the structure of these scales. For example, Onwuegbuzie et al. (2009) cautioned questionnaire designers about using midpoint or neutral categories in SET scales. Based on several studies, Onwuegbuzie and Weems (2004; Weems & Onwuegbuzie, 2001) argued that the inclusion of a midpoint option attenuates the internal consistency of SET scores. Sedlmeier (2006) suggested that the way in which rating scales are constructed may also have an impact on SET scores. Sedlmeier's study addresses the effects of three types of scales: (a) endpoint numbering (uni-polar vs. bipolar scales), (b) different ranges in scales, and (c) the ordering of choices. Two of these effects (endpoint numbering and different ranges in scales) are quite substantial and should therefore be considered when constructing SET questionnaires.

Robertson (2004) concluded that SET scores can be affected by item saliency and the position of questions in the questionnaire. Moreover, the SET scores observed in that study improved when students were asked to provide explanations for their answers. With regard to the number of response options, Landrum and Braitman (2008) reported that students use a greater range of points on a 5-point scale than they do on a 10-point scale. Students are more accurate using a 5-point scale, as it is easier to differentiate between five options than it is to distinguish between 10. In a study of response patterns in SET forms, Darby (2008) found that students tend to respond at the favorable end of evaluation scales, which does not mean that all courses were—in actual fact—good. For this reason, Darby argued that SET reports should include a means of comparison by, for instance, asking students to rank a course in comparison to other courses.

Recent SET studies have also focused on acquiescence (yea saying) as a response style, although the results have been mixed. Although a recent study (Spooren, Mortelmans, & Thijssen, 2012) yielded no evidence of acquiescence in SET scores, Richardson (2012) identified both acquiescence and extreme responding as consistent traits in SET. The precise impact of these traits remains unclear, but caution is advised with regard to possible bias due to acquiescence and extreme responding in SET results. To this end, studies by Dolnicar and Grun (2009) and Spooren et al. (2012) provide lists of recommendations for avoiding, controlling, and correcting for acquiescence and extreme responding in SET. These lists include using semibalanced scales, calculating reliability estimates, counting frequencies, comparing groups of students, and employing such correction methods as the subtraction of individual means and division by the individual standard deviations.

Acquiescence sometimes results from excessively demanding SET practices in many institutions, which overburden students with evaluations. Sampling therefore appears to be an efficient strategy that does not decrease the validity and reliability of the results (Kreiter & Laksham, 2005). Roszkowski and Soven (2010) argued against the use of balanced scales and advocated using only positively worded items in SET questionnaires. In their opinion, the use of bi-directional (i.e., positive and negative) item wording produces ambiguous results, due to carelessness on the part of students. Given that response patterns might also emerge from poor item wording (e.g., vague, unclear, too difficult, irrelevant), attention should be paid to the formulation of items. Based on think-aloud interviews with students, Billings-Gagliardi, Barrett, and Mazor (2004) observed that students understand educational terms in different ways and therefore make different judgments.

Although many studies have been conducted on the reliability, validity, and utility of scales, most SET forms also include open-ended questions. Students are invited to share more specific opinions and suggestions concerning both the course and the teacher. In an analysis of written comments from students, Nasser and Fresko (2009) observed that such comments are more often positive (59% positive units and 41% negative units) and general (rather than specific), and that they correlate with answers to the closed-ended questions in the questionnaire, as well as to specific characteristics of the course (correlations ranged between .23 and .57). The latter suggests that both closed-ended and open-ended questions should be included in SET forms, as written comments allow students to explain the scores that they assign for closed-ended items and to draw attention to topics that were not addressed in the closed-ended part of the form.

All of these findings suggest that questionnaire designers should be aware of the consequences of their choices (single item vs. Likert-type scale approach, the number of options, midpoint options) when constructing SET items, since their substantive validity is at risk. Response patterns, neutral responses, favorable answers, and different conceptions concerning educational terms might greatly influence SET scores.

Outcome validity. The previous sections demonstrate that SET scores may be affected by the instruments used, as well as by the opinions of student, perhaps even to the point of challenging their validity. Furthermore, much of the existing SET literature focuses on these topics. Even if all of these biasing challenges are under control, however, and even if SET provides valid information concerning the quality of teaching, it is still possible for such evaluations to be administered and used in inappropriate ways. Use affects the outcome validity of SET. Onwuegbuzie et al. (2009) argued that evidence concerning the outcome validity of SET may be the weakest of all evidence regarding validity issues.

Penny (2003) stated that the ways in which administrators engage with SET constitute one of the greatest threats to the validity of SET. Although guidelines for the collection and interpretation of SET data are available, many SET users are not sufficiently trained to handle these data, and they may even be unaware of their own ignorance. Moreover, they lack knowledge about the existing research literature on SET. Although the misuse and mis-collection of data might have consequences for both the improvement of teaching and the careers of the teachers involved, little research is available concerning this topic. In this section, we provide an overview of recent SET research concerning the collection and interpretation of SET data, which focuses primarily on attitudes toward SET and the relationship between SET and the improvement of teaching.

Students' attitudes toward SET. Students' attitudes toward the goals of SET are apparently important when collecting SET. If students see no connection between their efforts in completing SET questionnaires and the outcomes of these evaluations (e.g., teacher awards or improvements in teaching or course organization), such evaluations may become yet another routine task, thus leading to mindless evaluation behavior (Dunegan & Hrivnak, 2003). Spencer and Schmelkin (2002) reported from a mail survey to a random sample of students that students are generally willing to participate in SET procedures, and that they do not fear possible repercussions for giving negative evaluations (the mean scores on the factors Reluctance to Do Evaluations and Potential Repercussion Against Students, as measured by means of a 7-point scale with 1 as *disagree very strongly*, were 2.94 and 2.24, respectively). Nevertheless, they have little confidence that their evaluations are actually taken into account by either administrators or teachers (the mean scores on the factors Impact of Teaching on Students and Student Opinion Taken Seriously, as measured by means of a 7-point scale with 1 as *disagree very strongly*, were 4.55 and 4.28, respectively).

Students are also ambivalent about the relative utility of the SET process. Chen and Hoshower (2003) observed that, according to the students, providing feedback for the improvement of teaching is the most attractive outcome of a teaching-evaluation system. The expectations that students have concerning this outcome have a significant impact on their motivation to participate in evaluations. This is an important finding, as response rates in SET are generally low (fluctuating between 30% and 50%), especially in the case of online course evaluations (Arnold, 2009; Dommeyer, Baum, Hanna, & Chapman, 2004; Layne, Decristoforo, & McGinty, 1999), and might affect SET scores (McPherson, 2006).

With regard to their use of SET, students reported that they find SET somewhat useful (e.g., for course selection), although there is variation according to frequency of use, as well as according to student and program characteristics (Beran, Violato, Kline, & Frideres, 2009). Although students are more likely to choose courses that have good SET results (if they are available), the possibility of acquiring useful knowledge remains the most important selection criterion (Howell & Symbaluk, 2001; Wilhelm, 2004). Using SET for administrative decision-making was not found to be an important motivator for student participation in SET (Chen & Hoshower, 2003).

Teachers' attitudes toward SET. Teachers' attitudes toward SET are important for both the collection and the use of SET, given that the usefulness of these evaluations for the improvement of teaching depends upon the extent to which teachers respond to and use them (Ballantyne, Borthwick, & Packer, 2000). Nevertheless, Moore and Kuol (2005a) argued that surprisingly few studies examine faculty perceptions and the nature of teacher reaction to student feedback. Moore and Kuol (2005b) developed a tentative quadrant for understanding teacher reactions to SET (i.e., endorsement, ego protection, problem solving, and repair), based on a comparison of positive/negative self-evaluations with positive/negative SET. The authors observed two risks related to these reactions: fixation on minor issues (e.g., making changes to the layout of a PowerPoint presentation) and de-motivation, dejection, and withdrawal from the commitment to teaching effectiveness. Yao and Grady (2005) found from interviews with 10 faculty members that teachers care about feedback from students, although they experience anxiety and tension concerning the summative purposes of SET.

The ways in which teachers use SET varies according to background and experience. Arguing that responding to feedback is indeed a complex process, Arthur (2009) developed a typology of factors (e.g., personality, student characteristics, teaching and learning strategies) that affect teachers' individual responses to negative feedback (i.e., tame, blame, reframe, shame). Understanding the ways in which instructors respond to SET could help to overcome the doubts that teachers have regarding the validity of SET as an indicator of teaching quality, as well as their differing perceptions regarding the accuracy of SET (Simpson & Siguaw, 2000).

In general, teachers tend to agree that SET is an acceptable means of assessing institutional integrity and that it may be useful for administrative decision-making. Beran and Rokosh (2009) reported from a survey to 262 university teachers that 84% of the respondents support the use of SET in general, and that 62% of the respondents feel that department heads and deans make proper use of SET reports. Gender differences can be observed in perceptions of SET, however, with SET apparently having a greater negative impact on female teachers as they report a strong or moderate impact more often than male teachers when asked, “How much impact do you think your gender has on their evaluation of you?” (Kogan, Schoenfeld-Tacher, & Helleyer, 2010).

Based on interviews with 22 teachers, Burden (2008, 2010) observed a common recognition of the importance of SET. Nevertheless, only four of the teachers interviewed reported seeing the teacher feedback provided by SET as amounting to little more than hints and tips, as the evaluations did not reflect their perceptions of good teaching. The results of this study are supported by quantitative research results. Nasser and Fresko (2002) found from a survey with 101 instructors at a teacher’s college that instructors consider SET of little value for the improvement of their teaching, and that teachers make little or no use of student feedback. In the above-mentioned study, Beran and Rokosh (2009) found that SET results are used for improving general teaching quality (57%), for refining overall instruction (58%), and for improving lectures (54%). SET results are least often used for specific changes in particular courses, such as textbooks (23%), examinations (24%), student assignments (28%), support materials (34%), or for refining instructional objectives (40%).

Adminstrators’ attitudes toward SET. Although we are not aware of any recent study that include administrators’ attitudes toward SET, it is reasonable to expect that they would be more positive with regard to the use and validity of such evaluations, as they provide a quick and easy indicator of teaching performance (Sproule, 2000). Nevertheless, administrators have challenged the validity of SET based on limited psychometric knowledge (Franklin, 2001; Sproule, 2000; Wolfer & Johnson, 2003). Administrators prefer aggregated and overall measures of student satisfaction, often failing to consider both basic statistical and methodological matters (e.g., response rate, score distribution, sample size) when interpreting SET (Gray & Bergmann, 2003; Menges, 2000) and making spurious inferences based on these data. For example, Franklin (2001) reported that about half of the SET administrators involved in the study were unable to provide sound answers to several basic statistical questions. The proper collection and interpretation of SET data depend upon administrators having sound methodological training and regular briefing on the major findings and trends in the research field.

SET and the improvement of teaching. An important outcome of SET would be, as mentioned above, to provide student feedback for the improvement of teaching in particular courses. In the previous paragraph, we argued that many teachers do not find SET very helpful for such formative purposes and that they tend to ignore the comments and suggestions that students provide. These findings suggest that SET ultimately does not achieve the goal of providing useful information to an important stakeholder, with the ultimate goal of improvement. One important question addressed in the recent SET literature, therefore, involves the relationship between SET and the improvement of teaching. Davidovitch and Soen (2006b) showed that SET improves over time (with the age and seniority of teachers as particularly important predictors). Contrary to these results, however, a study by Kember et al. (2002) based on multiyear SET data from one university revealed no evidence that such evaluations contribute to the improvement of teaching, as SET scores did not increase over the years. These findings could be explained by several factors, including the organization and goals of SET in particular institutions, as well as the quality of the instruments and procedures that are used.

Consultative feedback on SET. Another possible explanation is that the student feedback obtained from the questionnaire is not used effectively. Marsh (2007a) concurred, saying that student feedback alone is not sufficient to achieve improvement in teaching. Using a multilevel growth-modeling approach, Marsh (2007a) demonstrated that SET reports are highly stable over time, including with regard to the individual differences between teachers. It is therefore important for teachers to have the opportunity to consult with colleagues or educational experts about their SET reports. In a longitudinal study, Dresel and Rindermann (2011) observed that consulting with faculty about their SET has a moderate to large positive effect (.68) on teaching quality, even when controlling for variables reflecting bias and unfairness. Lang and Kersting (2007) found that providing feedback by SET reports alone (without consultation) is far less effective than many assume in the long run. They noted a strong increase in SET results the next semester, which was followed by declines over the next three semesters.

Nasser and Fresko (2001) provided a typology of teachers who seek voluntary peer consultation regarding their SET reports. Three attributes were associated with this form of help seeking: lack of prior teacher training, teaching lecture courses, and being female. In addition, instructors were satisfied with their consultations, although they subsequently made few changes in their teaching. Relatedly, a meta-analysis by Penny and Coe (2004) on the effectiveness of consultation on student feedback showed that not all consultation practices are effective in improving teaching effectiveness. Consultative feedback should consist of more than simply interpreting the results and providing advice for teaching improvement. These authors listed eight strategies that are important when providing consultative feedback: (a) active involvement of teachers in the learning process, (b) use of multiple sources of information, (c) interaction with peers, (d) sufficient time for dialogue and interaction, (e) use of teacher self-ratings, (f) use of high-quality feedback information, (g) examination of conceptions of teaching, and (h) setting of improvement goals.

Predicting SET. Another strategy involves highlighting the discrepancy between predicted and actual ratings, which, according to Nasser and Fresko (2006), can serve as an impetus for teaching improvement. According to these authors, teachers are generally quite good at predicting their SET scores. Nevertheless, the results revealed a trend in which teachers with lower ratings tend to overestimate their SET, and those with higher ratings tend to underestimate their SET (effect sizes of significant differences based on *t* tests between teachers' predictions and SET results ranged between .61 and 1.30). It is clear that all of these strategies lead to the inclusion of SET in a more holistic approach that stimulates teachers to be and remain to be reflective practitioners concerning their teaching, instead of merely taking note of the next SET report.

Criterion-Related Validity

As mentioned above, SET research reveals moderate to large positive correlations between SET scores and other indicators of teaching quality (e.g., student achievement, alumni ratings, self-ratings). These coefficients provide strong evidence for the concurrent and predictive validity of SET instruments' scores. In recent years, however, electronic evaluation appears to have replaced the classic paper-and-pencil questionnaire as the most common means of gathering SET in institutions throughout the world (Arnold, 2009; Nulty, 2008). Recent research has examined the validity of SET results that are obtained from such electronic procedures to ascertain if these procedures provide SET scores that are comparable to those obtained from the more classic paper-and-pencil procedures. In this section, we discuss research results that focus on the relationship between paper-and-pencil SET procedures and electronic SET procedures. Second, we consider the rise of online SET platforms (such as RateMyProfessors) and their relationship with SET scores obtained from institutional procedures.

Concurrent validity of electronic versus paper-and-pencil SET procedures. The primary reasons given for shifting to electronic SET include the following: (a) greater accessibility to students, (b) quick and accurate feedback, (c) no disruption of class time, (d) more accurate analysis of the data, (e) better written comments, (f) guaranteed student anonymity (e.g., decreased risk of recognition due to handwriting), (g) decreased vulnerability to faculty influence, (h) lower costs, and (i) reduced time demands for administrators (Anderson, Cain, & Bird, 2005; Ballantyne, 2003; Bothell & Henderson, 2003; Bullock, 2003; Tucker, Jones, Straker, & Cole, 2003). Some parties nevertheless fear that SET results obtained in this way are easier to trace and can be consulted by almost everyone (Gamliel & Davidovitz, 2005).

Moreover, response rates in such evaluation procedures are lower than is the case with paper-and-pencil questionnaires (Gamliel & Davidovitz, 2005). Dommeyer et al. (2004) reported average response rates of 70% for in-class surveys and 29% for online surveys. Johnson (2003) suggested several strategies for increasing electronic SET response rates, including encouragement by the faculty (i.e., if faculty members show genuine interest in SET, students will be more motivated to participate) and increasing the intrinsic motivation of students to participate (e.g., by highlighting their important role as raters), providing access to the electronic evaluation system, and clear instructions concerning participation in the SET process.

Several studies have investigated whether the shift toward electronic evaluations has affected SET scores. Studies by Leung and Kember (2005) and by Liu (2006) revealed no significant differences between SET scores obtained from paper-and-pencil evaluations and those obtained through electronic evaluations. These results support the concurrent validity of both types of instruments, although Venette, Sellnow, and McIntyre (2010) reported that student comments in electronic evaluations are more detailed than are those in paper-and-pencil questionnaires. At the aggregate level, Barkhi and Williams (2010) noted that electronic SET scores are lower than are those obtained with paper-and-pencil surveys. These differences disappear, however, when controlling for course and instructor. Moreover, electronic SET instruments generate more extreme negative responses to Likert-type items than do paper-based surveys. Paper-and-pencil questionnaires have traditionally been administered during the last class of a particular course, thus making them subject to little or no influence from the examination for that course. In contrast, Arnold (2009) identified differences in SET scores obtained in electronic surveys, depending upon whether they were gathered before and after the examinations. These differences, however, applied only to students who had not passed the examinations. It is important to consider whether the period in which the surveys can be completed is scheduled to take place before or after the examinations.

In summary, the literature shows that electronic SET procedures perform as well as traditional paper-and-pencil evaluation forms do, and that they yield similar results. Although electronic surveys obviously offer considerable advantages, their greatest challenge continues to involve increasing the response rate.

Concurrent validity of online ratings of professors. In recent years, the territory of SET has expanded beyond the exclusive domain of institutions to the World Wide Web through such faculty-rating sites as RateMyProfessors.com, PassCollege.com, ProfessorPerformance.com, Ratingsonline.com, and Reviewum.com (Otto, Sanford, & Ross, 2008). The homepage of the most popular site, RateMyProfessors.com, states that, in 2011, the website counted more than 10 million completed rating forms for more than one million teachers in more than 6,500 (Anglo-Saxon) universities and colleges. The rating form consists of five single-item questions concerning the easiness, clarity, and helpfulness of the teacher, as well as the student's level of interest prior to attending class and the use of the textbook during the course. Students are also asked to provide other information, including the title of the course and their own course attendance and grade, and they have the opportunity to add additional detailed comments about the course or the professor. Finally, students are asked to rate the appearance of the teacher involved as "hot" or "not" (although the website suggests that this rating is "just for fun").

The RateMyProfessors.com website is subject to a noncontrolled self-selection bias (since we can assume that only those students who really liked or disliked a teacher will be more likely to register and to share their experiences via such environments), which has consequences for the representativeness, validity, and reliability of the results (for an overview, see Davison & Price, 2009). Data from these websites should therefore be taken with a grain of salt, and they should not be used for summative evaluations. Nevertheless, many students use these ratings as a source of information about their teachers (Otto et al., 2008). Researchers have recently begun studying the comments and ratings that are available on the RateMyProfessors website in order to learn more about their validity and their relationship to the more traditional forms of SET (as organized at the institutional level). Silva et al. (2008) found that the focus of ratings and comments on the website were very similar to those obtained through traditional evaluations, as they primarily concern teaching characteristics, personality, and global quality. Otto et al. (2008) observed that the online ratings on the RateMyProfessors website reflected student learning, thus possibly constituting a valid measure of teaching quality. In addition, there were no gender differences in the ratings. Besides, ratings on the RateMyProfessors website show statistically significant positive correlations (that exceed .60) with institutionally based SET (Sonntag, Bassett, & Snyder, 2009; Timmerman, 2008). In general, more lenient instructors receive higher overall quality ratings. Stuber, Watson, Carle, and Staggs (2009) observed that, controlling for other predictors, Instructor's Easiness predicted 50% of the variance in the scores on the Overall Quality measure. Timmerman (2008) found similar results and showed that this association can be partially explained by the fact that student's learning is associated with student conceptions of an instructor's perceived easiness.

As identified by Felton, Mitchell, and Stinson (2004), there is a positive correlation between overall ratings and the leniency and sexiness of instructors (correlations were .61 and .30, respectively). Finally, Freng and Webber (2009) find that the "hotness" variable accounted for almost 9% of the variance in SET scores on the RateMyProfessors website. This might strengthen the argument of those who found relationships between physical attractiveness and SET in institutional-based studies (see, e.g., the above mentioned studies by Feely, 2002; Gurung & Vespia, 2007). Still, Freng and Webber's noted that students rate a teacher's hotness on a dichotomous scale rather than a Likert-type scale, thus failing to capture a broader range of variability in attractiveness. The mixed results of these studies and many methodological concerns (self-selection bias, poorly designed questionnaires, the absence of data on the psychometric properties of the instrumentarium) suggest that student evaluations from these websites should be interpreted with great caution.

Discussion

As demonstrated in the previous sections, SET remains a current yet controversial topic in higher education as well as in education research. Many stakeholders are not convinced of the usefulness and validity of SET for both formative and summative purposes. Research on SET has thus far failed to provide clear answers to several critical aspects concerning the validity of SET. This article provides an overview of the recent research on the use and the validity of SET. In this final section, we summarize the most important findings of the present study. We relate these findings to the meta-validation framework for SET (Onwuegbuzie et al., 2009) and formulate several suggestions for further research in the field of SET.

Content-Related Validity

Although SET questionnaires can be assumed to have face validity (Onwuegbuzie et al., 2009), recent SET research has revealed differences in the perspectives that various stakeholders have of good teaching. Such differences threaten both the item validity and the sampling validity of SET instruments, as it is impossible to gather information concerning the extent to which SET instruments provide adequate and complete representations of particular content areas. The renewed call for a common conceptual framework with regard to effective teaching would offer questionnaire architects the opportunity to test their instruments in these areas of validity as well.

Construct-Related Validity

Structural validity. Our review has further shown that many SET instruments have been subjected to thorough validation procedures, although many of these procedures were conducted after the fact. Useful SET instruments are based on both educational theory and the rigorous investigation of their utility and validity (for examples, see, e.g., Marsh et al., 2009; Onwuegbuzie et al., 2007). Nevertheless, many *ad hoc* SET instruments that have never been tested continue to be used for administrative decision-making. When adapting existing instruments to other educational contexts, users are advised to be very cautious of the applicability paradigm (Marsh & Roche, 1997) and to test the validity of the instrument again in the new context. For example, Rindermann and Schofield (2001) demonstrated the validity and reliability of their instrument across six traditional and technical German universities.

It will also be important to test the long-term stability of SET instruments' scores that have been found valid. For example, because the didactic approaches in many institutions have shifted from teacher-centered toward student-centered, it might be quite important to retest existing SET instruments for their utility within these changed contexts—or to determine whether new instruments are needed. In a similar vein, we should consider the evaluation behavior of students when using the same SET instruments for many years. Repeated use might influence their responses in their "umpteenth" evaluation.

Convergent validity. There is no consensus regarding the strength of the correlation between SET and student achievement. This lack has much to do with the measure of learning (i.e., grades, students' perceptions of learning, test outcomes) that was used in the research literature on this topic. Clayson (2009) argued that the more objective the learning is measured, the lower the association between achievement and SET will be. Nevertheless, student achievement should not be measured solely by grades that students make or their perceptions of learning. For example, Clayson (2009) listed five alternatives for increasing the stringency of controls when mapping student learning: using class means (instead of individual means), using common tests in multiple section courses, conducting pretests and posttests, monitoring performance in future classes, and using standardized tests. In this regard as well, agreements are needed in order to determine student achievement (i.e., which measure(s) can be used to investigate the relationships between student achievement and SET scores).

Discriminant validity and divergent validity. The most prominent topic in the SET literature continues to involve the discriminant validity of SET, given the frequency with which new bias studies are published. Unfortunately, the sometimes-contradictory findings concerning the relationships (or strength of the relationships) between SET and the characteristics of students, courses, and teachers do not promote any conclusive idea of factors that could potentially bias SET scores. This issue is closely related to the number of control variables included in these studies, the way in which these variables are measured, the various research techniques applied, and the characteristics of the samples. It is very difficult to make valuable statements concerning the generalizability of the results (for instance, concerning global effects sizes of such a characteristic on SET scores), as these results are genuinely mixed based on strong and less strong findings on both sides. In addition, recent studies also address the question of whether personal traits and/or halo effects occur in SET, given the possibility that such evaluations could be influenced by psychodynamic aspects that may have consequences for the interpretation of the results.

Outcome validity. Recent research on the outcome validity of SET provides interesting results concerning the attitudes of both teachers and students toward the utility of SET, as well as their actual practices with regard to completing SET forms and the use of their results for the improvement of teaching. In general, students are willing to participate in SET procedures, although they think that teachers and institutions make little or no use of the results. Teachers agree with the use of SET for personnel decisions, as well as to demonstrate the quality of education at institutions, although they make little use of SET in order to improve their teaching. Moreover, responding to SET appears to be more difficult than many stakeholders may assume. It is therefore important for SET to be conducted with great caution and for teachers to count on peers, colleagues, and administrators when interpreting their SET results. Finally, it is important for SET administrators to be trained in both statistics and educational theory, in addition to being well informed about the SET literature. A skilled administrator can remove many of the concerns that teachers have with regard to SET.

The findings concerning SET and the long-term improvement of teaching suggest that such evaluations alone do not lead to better teaching. For this reason, (a) SET should be embedded within a more holistic approach to the evaluation of teaching, in which teachers make a serious effort to reflect upon the improvement of their teaching in a course; (b) teachers should be able to rely on expert consultation concerning their SET scores; and (c) SET should not be the sole means used to map a teacher's teaching (or progress therein).

Generalizability. When reviewing the literature, it becomes clear that most studies in the field suffer from two important limitations that confine their generalizability since, in general, it can be said that these studies were executed in *a particular setting* using *a particular instrument*. First, it is fair to say that most studies were done using nothing more or less than a well-designed (institutional) SET questionnaire, although some standardized questionnaires (such as SEEQ or CEQ) are widely available. Cross-validation procedures in other institutions are needed to demonstrate the generalizability of these institution-based instruments in other settings. Second, the results of many studies are influenced by the SET practice at the institutions. It is probable that most of the contradictory research results on SET are (at least partly) due to the great variety of methods, measures, controlling variables, SET instruments, and populations used in these studies.

Criterion-Related Validity

SET research reveals a positive correlation between SET scores and other indicators of teaching quality (e.g., student learning outcomes, alumni ratings, self-ratings). This supports the criterion-related validity of scores on SET instruments. Little is known, however, concerning whether the various well-validated SET instruments (e.g., the SEEQ or the CEQ) yield similar results when adopted in identical SET settings. Multitrait–multimethod analysis (in which these instruments are used as different measures of several dimensions of effective teaching) or, more simply, analysis of the correlations between the scores generated by the instruments could yield further evidence on the concurrent validity of these instruments.

Online SET has become the norm at many institutions of higher education. This development has understandably generated many studies on the validity of the results from Web-based student evaluations. For institutions, the results obtained with online SET instruments are similar to those obtained with paper-and-pencil instruments, although students provide more comments in an online environment. Low response rates constitute a major disadvantage of online SET, and this has consequences for the interpretation of the results (e.g., it is not clear whether they are representative of the entire population). It would be interesting to learn (a) which types of students participate in SET and which do not and (b) whether the perceptions of participants differ from those of nonparticipants. Researchers have found that internet-based SET systems yield results that are comparable to those obtained within the institutions. We nevertheless advise against relying on these websites, due to self-selection bias on the part of students, the psychometric properties of the instruments used, and the relationship between SET results and teacher characteristics that are unrelated to effective teaching (e.g., their hotness or sexiness).

Conclusion

This review of the state of the art in the literature has shown that the utility and validity ascribed to SET should continue to be called into question. Next to some, although much-researched, topics such as the dimensionality debate and the bias question, new research lines are delineated (i.e., the utility of online SET, teacher personal characteristics affecting SET). Our systematic use of the meta-validity framework of Onwuegbuzie et al. (2009), however, shows that many types of validity of SET remain at stake. Because conclusive evidence has not been found yet, such evaluations should be considered fragile, as important stakeholders (i.e., the subjects of evaluations and their educational performance) are often judged according to indicators of effective teaching (in some cases, a single indicator), the value of which continues to be contested in the research literature.

References

Abrami, P. C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. *New Directions for Teaching and Learning*, *43*, 97–111. doi:10.1002/tl.37219904309

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness-generalizability of 'N = 1' research. Comment on Marsh (1991). *Journal of Educational Psychology*, *83*, 411–415. doi:10.1037/0022-0663.83.3.411

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*, 153–166. doi:10.1023/A:1008168421283

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*, 431–441. doi:10.1037/0022-3514.64.3.431

*Anderson, H. M., Cain, J. C., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education*, *69*, 34–43. Retrieved from http://archive.ajpe.org/view.asp?art=aj690105&pdf=yes

*Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education*, *30*, 723–748. doi:10.1080/03075070500340101

*Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research*, *48*, 215–224. doi:10.1016/j.ijer.2009.10.001

*Arthur, L. (2009). From performativity to professionalism: Lecturer's responses to student feedback. *Teaching in Higher Education*, *14*, 441–454. doi:10.1080/13562510903050228

*Balam, E., & Shannon, D. (2010). Student ratings of college teaching: A comparison of faculty and their students. *Assessment and Evaluation in Higher Education*, *35*, 209–221. doi:10.1080/02602930902795901

*Ballantyne, C. (2003). Online evaluations of teaching: An examination of current practice and considerations for the future. *New Directions for Teaching and Learning*, *96*, 103–112. doi:10.1002/tl.127

*Ballantyne, R., Borthwick, J., & Packer, J. (2000). Beyond student evaluation of teaching: Identifying and addressing academic staff development needs. *Assessment and Evaluation in Higher Education*, *25*, 221–236. doi:10.1080/713611430

*Barnes, D., Engelland, B., Matherne, C., Martin, W., Orgeron, C., Ring, J., et al. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal*, *42*, 199-213.

*Barkhi, R., & Williams, P. (2010). The impact of electronic media on faculty evaluation. *Assessment and Evaluation in Higher Education*, *35*, 241–262. doi:10.1080/02602930902795927

*Barth, M. M. (2008). Deciphering student evaluations of teaching: A factor analysis approach. *Journal of Education for Business*, *84*, 40–46. doi:10.3200/JOEB.84.1.40-46

*Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education*, *18*, 91–106. doi:10.1007/s11092-006-9001-8

*Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, *30*, 25–35. doi:10.1111/j.1471-6402.2006.00259.x

*Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, *27*, 253–265. doi:10.1016/j.econedurev.2006.08.007

Beecham, R. (2009). Teaching quality and student satisfaction: Nexus or simulacrum? *London Review of Education*, *7*, 135–146. doi:10.1080/14748460902990336

*Beran, T. N., & Rokosh, J. L. (2009). Instructor's perspectives on the utility of student ratings of instruction. *Instructional Science*, *37*, 171–184. doi:10.1007/s11251-007-9045-2

*Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment and Evaluation in Higher Education*, *30*, 593–601. doi:10.1080/02602930500260688.

*Beran, T., Violato, C., Kline, D., & Frideres, J. (2009). What do students consider useful about student ratings? *Assessment and Evaluation in Higher Education*, *34*, 519–527. doi:10.1080/02602930802082228

*Billings-Gagliardi, S., Barrett, S. V., & Mazor, K. M. (2004). Interpreting course evaluation results: Insights from thinkaloud interviews with medical students. *Medical Education*, *38*, 1061–1070. doi:10.1111/j.1365-2929.2004.01953.x

Blackmore, J. (2009). Academic pedagogies, quality logics and performative universities: Evaluating teaching and what students want. *Studies in Higher Education*, *34*, 857–872. doi:10.1080/03075070902898664

Bolivar, A. (2000). Student teaching evaluations: Options and concerns. *Journal of Construction Education*, *5*, 20–29. Retrieved from http://www.ascjournal.ascweb.org/

*Bosshardt, W., & Watts, M. (2001). Comparing student and instructor evaluations of teaching. *Journal of Economic Education*, *32*, 3–17. doi:10.1080/00220480109595166

*Bothell, T. W., & Henderson, T. (2003). Do online ratings of instruction make sense? *New Directions for Teaching and Learning*, *96*, 69–79. doi:10.1002/tl.124

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

*Brockx, B., Spooren, P., & Mortelmans, D. (2011). Taking the "grading leniency" story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. *Educational Assessment, Evaluation and Accountability*, *23*, 289–306. doi:10.1007/s11092-011-9126-2

*Bullock, C. D. (2003). Online collection of midterm student feedback. *New Directions for Teaching and Learning*, *96*, 95–102. doi:10.1002/tl.126

*Burden, P. (2008). Does the end of semester evaluation forms represent teacher's views of teaching in a tertiary education context in Japan? *Teaching and Teacher Education*, *24*, 1463–1475. doi:10.1016/j.tate.2007.11.012

*Burden, P. (2010). Creating confusion or creative evaluation? The use of student evaluation of teaching surveys in Japanese tertiary education. *Educational Assessment, Evaluation and Accountability*, *22*, 97–117. doi:10.1007/s11092-010-9093-z

*Burdsal, C. A., & Bardo, J. W. (1986). Measuring student's perception of teaching: Dimensions of evaluation. *Educational and Psychological Measurement*, *46*, 63–79. doi:10.1177/0013164486461006

*Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education*, *33*, 567–576. doi:10.1080/02602930701699049

*Campbell, H., Gerdes, K., & Steiner, S. (2005). What's looks got to do with it? Instructor appearance and student evaluations of teaching. *Journal of Policy Analysis and Management*, *24*, 611–620. doi:10.1002/pam.20122

Cashin, W. E., & Perrin, P. B. (1978). *IDEA Technical Report No. 4. Description of IDEA Standard Form Data Base*. Manhattan, KS: Center for Faculty Evaluation and Development in Higher Education.

Centra, J. A. (1998). *Development of The Student Instructional Report II*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Products/283840.pdf

*Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*, 495–518. doi:10.1023/A:1025492407752

*Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, *71*, 17–33. Retrieved from www.jstor.org/stable/2649280

*Chen, Y., & Hoshower, L. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education*, *28*, 71–88. doi:10.1080/02602930301683

*Cheung, D. (2000). Evidence of a single second-order factor in student ratings of teaching. *Structural Equation Modeling*, *7*, 442–460. doi:10.1207/S15328007SEM0703_5

*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*, 16–30. doi:10.1177/0273475308324086

*Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education*, *28*, 149–160. doi:10.1177/0273475306288402

*Coffey, M., & Gibbs, G. (2001). The evaluation of the Student Evaluation of Educational Quality Questionnaire (SEEQ) in UK higher education. *Assessment & Evaluation in Higher Education*, *26*, 89–93. doi:10.1080/02602930020022318

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*, 281–309. doi:10.3102/00346543051003281

*Cohen, E. H. (2005). Student evaluations of course and teacher: Factor analysis and SSA approaches. *Assessment & Evaluation in Higher Education*, *30*, 123–136. doi:10.1080/0260293042000264233

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349–354. doi:10.1037/h0047358

Crumbley, L. C., Flinn, R. E., & Reichelt, K. J. (2010). What is ethical about grade inflation and coursework deflation? *Journal of Academic Ethics, 8*, 187–197. doi:10.1007/s10805-010-9117-9

*Darby, J. A. (2008). Course evaluations: A tendency to respond "favourably" on scales? *Quality Assurance in Education*, *16*, 7–18. doi:10.1108/09684880810848387

*Davidovitch, N., & Soen, D. (2006a). Class attendance and students' evaluation of their college instructors. *College Student Journal*, *40*, 691–703.

*Davidovitch, N., & Soen, D. (2006b). Using students' assessments to improve instructors' quality of teaching. *Journal of Further and Higher Education*, *30*, 351–376. doi:10.1080/03098770600965375

*Davison, E., & Price, J. (2009). How do we rate? An evaluation of online evaluations. *Assessment & Evaluation in Higher Education*, *34*, 51–65. doi:10.1080/02602930801895695

*Dee, K. C. (2007). Student perceptions of high course workloads are not associated with poor student evaluations of instructor performance. *Journal of Engineering Education*, *96*, 69–78. Retrieved from http://www.jee.org/2007/january/6.pdf

*Delucchi, M. (2000). Don't worry, be happy: Instructor likability, student perceptions of learning, and teacher ratings in upper-level sociology courses. *Teaching Sociology*, *28*, 220–231. Retrieved from www.jstor.org/stable/1318991

*Dolnicar, S., & Grun, B. (2009). Response style contamination of student evaluation data. *Journal of Marketing Education*, *31*, 160–172. doi:10.1177/0273475309335267

*Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*, 611–623. doi:10.1080/02602930410001689171

Douglas, J., & Douglas, A. (2006). Evaluating teaching quality. *Quality in Higher Education*, *12*, 3–13. doi:10.1080/13538320600685024

*Dresel, M., & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: A multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education*, *52*, 717–732. doi:10.1007/s11162-011-9214-7

*Dunegan, K. J., & Hrivnak, M. W. (2003). Characteristics of mindless teaching evaluations and the moderating effects of image compatibility. *Journal of Management Education*, *27*, 280–303. doi:10.1177/1052562903027003002

Edström, K. (2008). Doing course evaluation as if learning matters most. *Higher Education Research & Development*, *27*, 95–106. doi:10.1080/07294360701805234

Eiszler, C. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*, 483–501. doi:10.1023/A:1015579817194

Ellis, L., Burke, D., Lomire, P., & McCormack, D. (2003). Student grades and average ratings of instructional quality. The need for adjustment. *The Journal of Educational Research*, *97*, 35–40. doi:10.1080/00220670309596626

*Emery, C. R., Kramer, T. R., & Tian, R. (2003). Return to academic standards: A critique of students' evaluations of teaching effectiveness. *Quality Assurance in Education*, *11*, 37–47. doi:10.1108/09684880310462074

*Ewing, V. L., Stukas, A. A., & Sheehan, E. P. (2003). Student prejudice against male and lesbian lecturers. *The Journal of Social Psychology*, *143*, 569–579. doi:10.1080/00224540309598464

*Feeley, T. H. F. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education*, *51*, 225–236. doi:10.1080/03634520216519

Feldman, K. A. (1997). Identifying exemplary teachers and teaching. Evidence from student ratings. In R. Perry, & J. Smart (Eds.), *Effective teaching in higher education. Research and Practice* (pp. 368-395). New York: Agathon.

*Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment and Evaluation in Higher Education*, *29*, 91–108. doi:10.1080/0260293032000158180

*Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning*, *87*, 85–100. doi:10.1002/tl.10001

*Freng, S., & Webber, D. (2009). Turning up the heat on online teaching evaluations: Does "hotness" matter? *Teaching of Psychology*, *36*, 189–193. doi:10.1080/00986280902959739

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

*Gamliel, E., & Davidovitz, L. (2005). Online versus traditional teaching evaluations: Mode can matter. *Assessment & Evaluation in Higher Education*, *30*, 581–592. doi:10.1080/02602930500260647

*Ghedin, E., & Aquario, D. (2008). Moving towards multidimensional evaluation of teaching in higher education: A study across four faculties. *Higher Education*, *56*, 583–597. doi:10.1007/s10734-008-9112-x

*Ginns, P., & Barrie, S. (2004). Reliability of single-item ratings of quality in higher education: A replication. *Psychology Reports*, *95*, 1023–1030. doi:10.2466/pr0.95.3.1023-1030

*Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education*, *32*, 603–615. doi:10.1080/03075070701573773

*Goldstein, G. S., & Benassi, V. A. (2006). Students' and instructors' beliefs about excellent lecturers and discussion leaders. *Research in Higher Education*, *47*, 685–707. doi:10.1007/s11162-006-9011-x

*Gray, M., & Bergmann, B. R. (2003). Student teaching evaluations: Inaccurate, demeaning, misused. *Academe, 89*, 44–46.

*Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality—Analysis of relevant factors based on empirical research. *Assessment & Evaluation in Higher Education*, *28*, 229–238. doi:10.1080/0260293032000059595

*Griffin, B. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology*, *26*, 534–552. doi:10.1006/ceps.2000.1075

*Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology*, *29*, 410–425. doi:10.1016/j.cedpsych.2003.11.001

*Guinn, B., & Vincent, V. (2006). The influence of grades on teaching effectiveness ratings at a Hispanic-serving institution. *Journal of Hispanic Higher Education*, *5*, 313–321. doi:10.1177/1538192706291138

*Gump, S. E. (2007). Student evaluation of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly*, *30*, 55–68.

*Gursoy, D., & Umbreit, W. T. (2005). Exploring students' evaluations of teaching effectiveness: What factors are important? *Journal of Hospitality and Tourism Research*, *29*, 91–109. doi:10.1177/1096348004268197

*Gurung, R., & Vespia, K. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology*, *34*, 5–10. doi:10.1080/00986280709336641

*Haladyna, T., & Amrein-Beardsley, A. (2009). Validation of a research-based student survey of instruction in a college of education. *Educational Assessment, Evaluation and Accountability*, *21*, 255–276. doi:10.1007/s11092-008-9065-8

*Hamermesch, D. S., & Parker, A. (2005). Beauty in the classroom: Instructor's pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*, 369–376. doi:10.1016/j.econedurev.2004.07.013

*Harnish, R. J., & Bridges, K. R. (2011). Effect of syllabus tone: Students' perceptions of instructor and course. *Social Psychology of Education*, *14*, 319–330. doi:10.1007/s11218-011-9152-4

*Harrison, P., Douglas, D., & Burdsal, C. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, *45*, 311–323. doi:10.1023/B:RIHE.0000019592.78752.da

*Heckert, T. M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness and student evaluations of teaching: Is it possible to "buy" better evaluations through lenient grading? *College Student Journal*, *40*, 588–596.

*Howell, A. J., & Symbaluk, D. G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology*, *93*, 790–796. doi:10.1037/0022-0663.93.4.790

*Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *Journal of Economic Education*, *36*, 29–42. doi:10.3200/JECE.36.1.29-42

Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of student's perceptions of teaching effectiveness. *Educational and Psychological Measurement*, *59*, 580–596. doi:10.1177/00131649921970035

Jauhiainen, A., Jauhiainen, A., & Laiho, A. (2009). The dilemmas of the "efficiency university" policy and the everyday life of university teachers. *Teaching in Higher Education*, *14*, 417–428. doi:10.1080/13562510903050186

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education*, *5*, 419–434. doi:10.1080/713699176

*Johnson, T. D. (2003). Online student ratings: Will students respond? *New Directions for Teaching and Learning*, *96*, 49–59. doi:10.1002/tl.122

*Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the Teacher Behavior Checklist. *Teaching of Psychology*, *37*, 16–20. doi:10.1080/00986280903426282

*Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology*, *33*, 84–90. doi:10.1207/s15328023top3302_1

*Kember, D., Jenkins, W., & Kwok, C.N. (2004). Adult students' perceptions of good teaching as a function of their conceptions of learning—Part 2. Implications for the evaluation of teaching. *Studies in Continuing Education*, *26*, 81–97. doi:10.1080/158037042000199461

*Kember, D., & Leung, D. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education*, *33*, 341–353. doi:10.1080/02602930701563070

*Kember, D., & Leung, D. (2011). Disciplinary differences in student ratings of teaching quality. *Research in Higher Education*, *52*, 279–299. doi:10.1007/s11162-010-9194-z

*Kember, D., Leung, D., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education*, *27*, 411–425. doi:10.1080/0260293022000009294

*Kember, D., & Wong, A. (2000). Implications for evaluation from a study of students' perceptions of good and poor teaching. *Higher Education*, *40*, 69–97. doi:10.1023/A:1004068500314

*Kim, C., Damewood, E., & Hodge, N. (2000). Professor attitude: Its effect on teaching evaluations. *Journal of Management Education*, *24*, 458–473. doi:10.1177/105256290002400405

Knapper, C. (2001). Broadening our approach to teaching evaluation. *New Directions for Teaching and Learning*, *88*, 3–9. doi:10.1002/tl.32

*Kogan, L., Schoenfeld-Tacher, R., & Helleyer, P. (2010). Student evaluations of teaching: Perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education*, *15*, 623–636. doi:10.1080/13562517.2010.491911

*Kohn, J., & Hartfield, L. (2006). The role of gender in teaching effectiveness ratings of faculty. *Academy of Educational Leadership Journal*, *10*, 121–137.

*Kreiter, C. D., & Laksham, V. (2005). Investigating the use of sampling for maximising the efficiency of student-generated faculty teaching evaluations. *Medical Education*, *39*, 171–175. doi:10.1111/j.1365-2929.2004.02066.x

Kulik, J. A. (2001). Student ratings: Validity, utility and controversy. *New Directions for Institutional Research*, *27*, 9–25. doi:10.1002/ir.1

*Landrum, R. E., & Braitman, K. A. (2008). The effect of decreasing response options on students' evaluation of instruction. *College Teaching*, *56*, 215–217. doi:10.3200/CTCH.56.4.215-218

*Lang, J. W. B., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science*, *35*, 187–205. doi:10.1007/s11251-006-9006-1

*Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review*, *27*, 417–428. doi:10.1016/j.econedurev.2006.12.003

Larsen, M. A. (2005). A critical analysis of teacher evaluation policy trends. *Australian Journal of Education*, *49*, 292–305.

Lattuca, L., & Domagal-Goldman, J. (2007). Using qualitative methods to assess teaching effectiveness. *New Directions for Institutional Research*, *136*, 81–93. doi:0.1002/ir.233

*Layne, B. H., Decristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, *40*, 221–232. doi:10.1023/A:1018738731032

*Leung, D. Y. P., & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the internet. *Research in Higher Education*, *46*, 571–591. doi:10.1007/s11162-005-3365-3

*Liu, Y. (2006). A comparison study of online versus traditional student evaluation of instruction. *International Journal of Instructional Technology and Distance Learning*, *4*, 15–29. Retrieved from http://www.itdl.org/

*Marks, R. B. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education*, *22*, 108–119. doi:10.1177/0273475300222005

Marsh, H. W. (1982). SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77–95. doi:10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology*, *76*, 707–754. doi:10.1037/0022-0663.76.5.707

Marsh, H. W. (1987). Student's evaluations of university teaching: Research findings, methodological issues, and directions for further research. *International Journal of Educational Research*, *11*, 253–388. doi:10.1016/0883-0355(87)90001-2

Marsh, H. W. (1991a). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami & d'Apollonia (1991). *Journal of Educational Psychology*, *83*, 416–421. doi:10.1037/0022-0663.83.3.416

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*, 285–296. doi:10.1037/0022-0663.83.2.285

*Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal*, *38*, 183–212. doi:10.3102/00028312038001183

*Marsh, H. W. (2007a). Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years. *Journal of Educational Psychology*, *99*, 775–790. doi:10.1037/0022-0663.99.4.775

*Marsh, H. W. (2007b). 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). New York: Springer.

Marsh, H. W., & Hovecar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education*, *7*, 9–18. doi:10.1016/0742-051X(91)90054-S

*Marsh, H. W., Muthèn, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439–476. doi:10.1080/10705510903008220

Marsh, H. W., & Roche, L A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility. *American Psychologist*, *52*, 1187–1197. doi:10.1037/0003-066X.52.11.1187

*Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology*, *92*, 202–228. doi:10.1037/0022-0663.92.1.202

*Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology*, *33*, 176–179. doi:10.1207/s15328023top3303_4

McKone, K. E. (1999). Analysis of student feedback improves instructor effectiveness. *Journal of Management Education*, *23*, 396–415. doi:10.1177/105256299902300406

*McPherson, M. A. (2006). Determinants of how students evaluate teachers. *Journal of Economic Education*, *37*, 3–20. doi:10.3200/JECE.37.1.3-20

*McPherson, M. A., & Todd Jewell, R. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, *88*, 868–881. doi:10.1111/j.1540-6237.2007.00487.x

*McPherson, M. A., Todd Jewell, R., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal*, *35*, 37–51. doi:10.1057/palgrave.eej.9050042

*Menges, R. J. (2000). Shortcomings of research on evaluating and improving teaching in higher education. *New Directions for Teaching and Learning*, *83*, 5–11. doi:10.1002/tl.8301

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003-066X.50.9.741

*Mohanty, G., Gretes, J., Flowers, C., Algozzine, B., & Spooner, F. (2005). Multimethod evaluation of instruction in engineering classes. *Journal of Personnel Evaluation in Higher Education*, *18*, 139–151. doi:10.1007/s11092-006-9006-3

Molesworth, M., Nixon, E., & Scullion, R. (2009). Having, being and higher education: The marketisation of the university and the transformation of the student into consumer. *Teaching in Higher Education*, *14*, 277–287. doi:10.1080/13562510902898841

*Montoneri, B., Lee, C. C., Lin, T. T., & Huang, S. L. (2011). A learning performance evaluation with benchmarking concept for English writing courses. *Expert Systems with Applications*, *38*, 14542–14549. doi:10.1016/j.eswa.2011.05.029

*Montoneri, B., Lin, T. T., Lee, C. C., & Huang, S. L. (2012). Application of data envelopment analysis on the indicators contributing to learning and teaching performance. *Teaching and Teacher Education*, *28*, 382–395. doi:10.1016/j.tate.2011.11.006

*Moore, S., & Kuol, N. (2005a). Students evaluating teachers: Exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education*, *10*, 57–73. doi:10.1080/1356251052000305534

*Moore, S., & Kuol, N. (2005b). A punitive bureaucratic tool or a valuable resource? Using student evaluations to enhance your teaching. In G. O'Neill, S. Moore, & B. McMullin (Eds.), *Emerging issues in the practice of university learning and teaching. Part 3: Developing and growing as a university teacher* (pp. 141–146). Dublin, Ireland: University of Limerick.

*Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37-questionnaire for student evaluations of teaching. *Educational Studies*, *35*, 547–552. doi:10.1080/03055690902880299

*Nasser, F., & Fresko, B. (2001). Interpreting student ratings: Consultation, instructional modification, and attitudes towards course evaluation. *Studies in Educational Evaluation*, *27*, 291–305. doi:10.1016/S0191-491X(01)00031-1

*Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, *27*, 187–198. doi:10.1080/02602930220128751

*Nasser, F., & Fresko, B. (2006). Predicting student ratings: The relationship between actual student ratings and instructor's predictions. *Assessment & Evaluation in Higher Education*, *31*, 1–18. doi:10.1080/02602930500262338

*Nasser, F., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation*, *35*, 37–44. doi:10.1016/j.stueduc.2009.01.002

*Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*, 301–314. doi:10.1080/02602930701293231

*Ogier, J. (2005). Evaluating the effect of a lecturer's language background on a student rating of teaching form. *Assessment & Evaluation in Higher Education*, *30*, 477–488. doi:10.1080/02602930500186941

Oleinik, A. (2009). Does education corrupt? Theories of grade inflation. *Educational Research Review*, *4*, 156–164. doi:10.1016/j.edurev.2009.03.001

*Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology*, *26*, 382–399. doi:10.1006/ceps.2000.1070

Olivares, O. J. (2003). A conceptual and analytic critique of student ratings of teachers in the USA with implications for teacher effectiveness and student learning. *Teaching in Higher Education*, *8*, 233–245. doi:10.1080/1356251032000052465

*Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity*, *43*, 197–209. doi:10.1007/s11135-007-9112-4

Onwuegbuzie, A. J., & Weems, G. H. (2004). Response categories on rating scales: Characteristics of item respondents who frequently utilize midpoint. *Research in the Schools*, *9*, 73-90.

*Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, *44*, 113–160. doi:10.3102/0002831206298169

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning*, *87*, 3–15. doi:10.1002/tl.23

*Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research*, *109*, 27–44. doi:10.1002/ir.2

*Otto, J., Sanford, D. A., & Ross, D. N. (2008). Does RateMyProfessor.com really rate my professor? *Assessment & Evaluation in Higher Education*, *33*, 355–368. doi:10.1080/02602930701293405

*Pan, D., Tan, G. S. H., Ragupathi, K., Booluck, K., Roop, R., & Ip, Y. K. (2009). Profiling teacher/teaching using descriptors derived from qualitative feedback: Formative and summative applications. *Research in Higher Education*, *50*, 73–100. doi:10.1007/s11162-008-9109-4

*Paswan, A. K., & Young, J. A. (2002). Student evaluation of instructor: A nomological investigation using structural equation modelling. *Journal of Marketing Education*, *24*, 193–202. doi:10.1177/0273475302238042

*Patrick, C. L. (2011). Student evaluations of teaching: Effects of the Big Five personality traits, grades and the validity hypothesis. *Assessment & Evaluation in Higher Education*, *36*, 239–249. doi:10.1080/02602930903308258

Paulsen, M. B. (2002). Evaluating teaching performance. *New Directions for Institutional Research*, *114*, 5–18. doi:10.1002/ir.42

*Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education*, *8*, 399–411. doi:10.1080/13562510309396

*Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research*, *74*, 215–253. doi:10.3102/00346543074002215

Platt, M. (1993). What student evaluations teach. *Perspectives on Political Science*, *22*, 29–40. doi:10.1080/10457097.1993.9944516

*Pozo-Munoz, C., Rebolloso-Pacheco, E., & Fernandez-Ramirez, B. (2000). The "Ideal Teacher". Implications for student evaluations of teaching effectiveness. *Assessment & Evaluation in Higher Education*, *25*, 253–263. doi:10.1080/02602930050135121

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education*, *16*, 129–150. doi:10.1080/03075079112331382944

Redding, R. (1998). Students' evaluations of teaching fuel grade inflation. *American Psychologist*, *53*, 1227–1228. doi:10.1037/0003-066X.53.11.1227

*Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal*, *34*, 91–115. doi:10.1080/01411920701492043

Remmers, H., & Brandenburg, G. (1927). Experimental data on the Purdue Rating Scale for Instruction. *Educational Administration and Supervision*, *13*, 519–527.

*Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education*, *46*, 929–953. doi:10.1007/s11162-005-6934-6

*Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education*, *30*, 387–415. doi:10.1080/02602930500099193

*Richardson, J. T. E. (2012). The role of response biases in the relationship between students' perceptions of their courses and their approaches to studying in higher education. *British Educational Research Journal*, *38*, 399–418. doi:10.1080/01411926.2010.548857

*Rindermann, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university teaching across courses and teachers. *Research in Higher Education*, *42*, 377–399. doi:10.1023/A:1011050724796

*Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology*, *133*, 19–35. doi:10.3200/GENP.133.1.19-35

*Robertson, S. I. (2004). Student perceptions of student perception of module questionnaires: Questionnaire completion as problem solving. *Assessment and Evaluation in Higher Education*, *29*, 663–679. doi:10.1080/0260293042000227218

*Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science*, *28*, 439–468. doi:10.1023/A:1026576404113

*Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negative worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*, 117–134. doi:10.1080/02602930802618344

*Santhanam, E., & Hicks, O. (2001). Disciplinary, gender and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education*, *7*, 17–31. doi:10.1080/13562510120100364

*Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction*, *16*, 401–415. doi:10.1016/j.learninstruc.2006.09.002

Seldin, P. (1993). The use and abuse of student ratings of professors. *Chronicle of Higher Education*, *39*, A40.

*Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education*, *25*, 397–405. doi:10.1080/713611436

*Silva, K. M., Silva, F. J., Quinn, M. A., Draper, J. N., Cover, K. R., & Munoff, A .A. (2008). Rate my Professor: Online evaluations of psychology instructors. *Teaching of Psychology*, *35*, 71–80. doi:10.1080/00986280801978434

*Simpson, P., & Siguaw, J. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, *22*, 199–213. doi:10.1177/0273475300223004

*Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication*, *30*, 64–77. doi:10.1080/07491409.2007.10162505

*Sonntag, M. E., Bassett, J. F., & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com. *Assessment & Evaluation in Higher Education*, *34*, 499–504. doi:10.1080/02602930802079463

*Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education*, *27*, 397–409. doi:10.1080/0260293022000009285

*Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation*, *36*, 121–131. doi:10.1016/j.stueduc.2011.02.001

*Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education. Development of an instrument based on 10 Likert scales. *Assessment and Evaluation in Higher Education*, *32*, 667–679. doi:10.1080/02602930601117191

*Spooren, P., Mortelmans, D., & Thijssen, P. (2012). Content vs. style. Acquiescence in student evaluations of teaching? *British Educational Research Journal*, *38*, 3–21. doi:10.1080/01411926.2010.523453

*Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives*, *8*, 50.

*Stack, S. (2003). Research productivity and student evaluation of teaching in social science classes. *Research in Higher Education*, *44*, 539–556. doi:10.1023/A:1025439224590

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

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

*Stodnick, M., & Rogers, P. (2008). Using SERVQUAL to measure the quality of the classroom experience. *Decisions Sciences Journal of Innovative Education*, *6*, 115–133. doi:10.1111/j.1540-4609.2007.00162.x

*Stuber, J. M., Watson, A., Carle, A., & Staggs, K. (2009). Gender expectations and on-line evaluations of teaching: Evidence from RateMyProfessors.com. *Teaching in Higher Education*, *14*, 387–399. doi:10.1080/13562510903050137

*Timmerman, T. (2008). On the Validity of RateMyProfessors.com. *Journal of Education for Business*, *84*, 55–61. doi:10.3200/JOEB.84.1.55-61

*Ting, K. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education*, *41*, 637–661. doi:10.1023/A:1007075516271

Titus, J. (2008). Student ratings in a consumerist academy: Leveraging pedagogical control and authority. *Sociological Perspectives*, *51*, 397–422. doi:10.1525/sop.2008.51.2.397

*Toland, M., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*, 272–296. doi:10.1177/001316440426866

*Tom, G., Tong, S. T., & Hesse, C. (2010). Thick slice and thin slice teaching evaluations. *Social Psychology of Education*, *13*, 129–136. doi:10.1007/s11218-009-9101-7

*Tucker, B., Jones, S., Straker, L., & Cole, J. (2003). Course evaluation on the web: Facilitating student and teacher reflection to improve learning. *New Directions for Teaching and Learning*, *96*, 81–94. doi:10.1002/tl.125

Valsan, C., & Sproule, R. (2005). The invisible hands behind the student evaluation of teaching: The rise of the new managerial elite in the governance of higher education. *Journal of Economic Issues*, *42*, 939–958.

*Venette, S., Sellnow, D., & McIntyre, K. (2010). Charting new territory: Assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education*, *35*, 101–115. doi:10.1080/02602930802618336

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education*, *23*, 191–210. doi:10.1080/0260293980230207

Weems, G. H., & Onwuegbuzie, A. J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development*, *34*, 166–176.

*Wendorf, C. A., & Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemporary Educational Psychology*, *30*, 190–206. doi:10.1016/j.cedpsych.2004.07.003

*Wilhelm, W. B. (2004). The relative influence of published teaching evaluations and other instructor attributes on course choice. *Journal of Marketing Education*, *26*, 17–30. doi:10.1177/0273475303258276

*Wolfer, T., & Johnson, M. (2003). Re-evaluating student evaluation of teaching: The teaching evaluation form. *Journal of Social Work Education*, *39*, 111–121.

*Yao, Y., & Grady, M. (2005). How do faculty make formative use of student evaluation feedback? A multiple case study. *Journal of Personnel Evaluation in Education*, *18*, 107–126. doi:10.1007/s11092-006-9000-9

Zabaleta, F. (2007). The use and misuse of student evaluation of teaching. *Teaching in Higher Education*, *12*, 55–76. doi:10.1080/13562510601102131

Authors

PIETER SPOOREN holds master's degrees in Educational Sciences and Quantative Analysis in the Social Sciences and a PhD in Social Sciences. He is affiliated as an educational advisor at the Faculty of Political and Social Sciences of the University of Antwerp (Belgium). His particular activities are educational innovation and evaluation of the educational process and of educators. His main research interests focus on students' evaluation of teaching (SET), in particular their use and validity.

BERT BROCKX holds a master's degree in Educational Sciences. He is affiliated as a predoctoral researcher at the Faculty of Political and Social Sciences of the University of Antwerp (Belgium). His main research interests focus on the validity of students' evaluation of teaching (SET).

DIMITRI MORTELMANS is an associate professor at the University of Antwerp. He is head of the Research Center for Longitudinal and Life Course Studies (CELLO). He publishes in the domain of family sociology and sociology of labor. Important topics of his expertise are ageing, divorce, and gender differences in career trajectories.

Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness

Anne Boring*

OFCE, SciencesPo, Paris

PSL, Université Paris-Dauphine, LEDa, UMR DIAL

Kellie Ottoboni and Philip B. Stark

Department of Statistics

University of California, Berkeley

January 5, 2016

*This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 612413, for the EGERA (Effective Gender Equality in Research and the Academia) European project.

The truth will set you free, but first it will piss you off.

Gloria Steinem

Abstract

Student evaluations of teaching (SET) are widely used in academic personnel decisions as a measure of teaching effectiveness. We show:

- SET are biased against female instructors by an amount that is large and statistically significant
- the bias affects how students rate even putatively objective aspects of teaching, such as how promptly assignments are graded
- the bias varies by discipline and by student gender, among other things
- it is not possible to adjust for the bias, because it depends on so many factors
- SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness
- gender biases can be large enough to cause more effective instructors to get lower SET than less effective instructors

These findings are based on nonparametric statistical tests applied to two datasets: 23,001 SET of 379 instructors by 4,423 students in six mandatory first-year courses in a five-year natural experiment at a French university, and 43 SET for four sections of an online course in a randomized, controlled, blind experiment at a US university.

1 Background

Student evaluations of teaching (SET) are used widely in decisions about hiring, promoting, and firing instructors. Measuring teaching effectiveness is difficult—for students, faculty, and administrators alike. Universities generally treat SET as if they primarily measure teaching effectiveness or teaching quality. While it may seem natural to think that students' answers to questions like "how effective was the instructor?" measure teaching effectiveness, it is not a foregone conclusion that they do. Indeed, the best evidence so far shows that they do not: they have *biases*[1] that are stronger than any connection they might have with effectiveness. Worse, in some circumstances the association between SET and an objective measure of teaching effectiveness is *negative*, as our results below reinforce.

Randomized experiments [Carrell and West, 2010, Braga et al., 2014] have shown that students confuse grades and grade expectations with the long-term value of a course and that SET are not associated with student performance in follow-on courses, a proxy for teaching effectiveness. On the whole, high SET seem to be a reward students give instructors who make them anticipate getting a good grade, for whatever reason; for extensive discussion, see Johnson [2003, Chapters 3–5].

Gender matters too. Boring [2015a] finds that SET are affected by gender biases and stereotypes. Male first-year undergraduate students give more *excellent* scores to male instructors, even though there is no difference between the academic performance of male students of male and of female instructors. Experimental work by MacNell et al. [2014] finds that when students think an instructor is female, students rate the instructor lower on every aspect of teaching, including putatively objective

[1] Centra and Gaubatz [2000, p.17] define bias to occur when "a teacher or course characteristic affects teacher evaluations, either positively or negatively, but is unrelated to criteria of good teaching, such as increased student learning."

measures such as the timeliness with which instructors return assignments.

Here, we apply nonparametric permutation tests to data from Boring [2015a] and MacNell et al. [2014] to investigate whether SET primarily measure teaching effectiveness or biases using a higher level of statistical rigor. The two main sources of bias we study are students' grade expectations and the gender of the instructor. We also investigate variations in bias by discipline and by student gender.

Permutation tests allow us to avoid contrived, counterfactual assumptions about parametric generative models for the data, which regression-based methods (including ordinary linear regression, mixed effects models, logistic regression, etc.) and methods such as t-tests and ANOVA generally require. The null hypotheses for our tests are that some characteristic—e.g., instructor gender—amounts to an arbitrary label and might as well have been assigned at random.

We work with course-level summaries to match how institutions use SET: typically, SET are averaged for each offering of a course, and those averages are compared across instances of the course, across courses in a department, across instructors, and across departments. Stark and Freishtat [2014] discuss statistical problems with this reduction to and reliance upon averages.

We find that the association between SET and an objective measure of teaching effectiveness, performance on the anonymously graded final, is weak and—for these data—generally not statistically significant. In contrast, the association between SET and (perceived) instructor gender is large and statistically significant: instructors whom (students believe) are male receive significantly higher average SET.

In the French data, *male* students tend to rate male instructors higher than they rate female instructors, with little difference in ratings by female students. In the US data, *female* students tend to rate (perceived) male instructors higher than they rate (perceived) female instructors, with little difference in ratings by male students.

The French data also show that gender biases vary by course topic, and that SET have a strong positive association with students' grade expectations.

We therefore conclude that SET primarily do not measure teaching effectiveness; that they are strongly and non-uniformly biased by factors including the genders of the instructor and student; that they disadvantage female instructors; and that it is impossible to adjust for these biases. SET should not be relied upon as a measure of teaching effectiveness. Relying on SET for personnel decisions has disparate impact by gender, in general.

2 Data

2.1 French Natural Experiment

These data, collected between 2008 and 2013, are a census of 23,001 SET from 4,423 first-year students at a French university students (57% women) in 1,177 sections, taught by 379 instructors (34% women). The data are not public, owing to French restrictions on human subjects data. Boring [2015a] describes the data in detail. Key features include:

- All first-year students take the same six mandatory courses: History, Macroeconomics, Microeconomics, Political Institutions, Political Science, and Sociology. Each course has one (male) professor who delivers the lectures to groups of approximately 900 students. Courses have sections of 10–24 students. Those sections are taught by a variety of instructors, male and female. The instructors have considerable pedagogical freedom.

- Students enroll in "triads" of sections of these courses (three courses per semester). The enrollment process does not allow students to select individual instructors.

The assignment of instructors to sets of students is as if at random, forming a *natural experiment*. It is reasonable to treat the assignment as if it is independent across courses.

- Section instructors assign interim grades during the semester. Interim grades are known to the students before the students submit SET. Interim grades are thus a proxy for students' grade expectations.

- Final exams are written by the course professor, not the section instructors. Students in all sections of a course in a given year take the same final. Final exams are graded anonymously, except in Political Institutions, which we therefore omit from analyses involving final exam scores. To the extent that the final exam measures appropriate learning outcomes, performance on the final is a measure of the effectiveness of an instructor: in a given course in a given year, students of more effective instructors should do better on the final exam, on average, than students of less effective instructors.

- SET are mandatory: response rates are nearly 100%.

SET include closed-ended and open-ended questions. The item that attracts the most attention, especially from the administration, is the *overall score*, which is treated as a summary of the other items. The SET data include students' individual evaluations of section instructors in microeconomics, history, political institutions, and macroeconomics for the five academic years 2008–2013, and for political science and sociology for the three academic years 2010–2013 (these two subjects were introduced in 2010). The SET are anonymous to the instructors, who have access to SET only after all grades have been officially recorded.

Table 1: Summary statistics of sections

course	# sections	# instructors	% Female instructors
Overall	**1,194**	**379**	**33.8%**
History	230	72	30.6%
Political Institutions	229	65	20.0%
Microeconomics	230	96	38.5%
Macroeconomics	230	93	34.4%
Political Science	137	49	32.7%
Sociology	138	56	46.4%

Data for a section of Political Institutions that had an experimental online format are omitted. Political Science and Sociology originally were not in the triad system; students were randomly assigned by the administration to different sections.

2.2 US Randomized Experiment

These data, described in detail by MacNell et al. [2014], are available at `http://n2t.net/ark:/b6078/d1mw2k`. Students in an online course were randomized into six sections of about a dozen students each, two taught by the primary professor, two taught by a female graduate teaching assistant (TA), and two taught by a male TA. In one of the two sections taught by each TA, the TA used her or his true name; in the other, she or he used the other TA's identity. Thus, in two sections, the students were led to believe they were being taught by a woman and in two they were led to believe they were being taught by a man. Students had no direct contact with TAs: the primary interactions were through online discussion boards. The TA credentials presented to the students were comparable; the TAs covered the same material; and assignments were returned at the same time in all sections (hence, objectively, the TAs returned assignments equally promptly in all four sections).

SET included an overall score and questions relating to professionalism, respectfulness, care, enthusiasm, communication, helpfulness, feedback, promptness, consistency, fairness, responsiveness, praise, knowledge, and clarity. Fourty-seven students in the four sections taught by TAs finished the class, of whom 43 submitted SET.

The SET data include the genders and birth years of the students;[2]the grade data do not. The SET data are not linked to the grade data.

3 Methods

Previous analyses of these data relied on parametric tests based on null hypotheses that do not match the experimental design. For example, the tests assumed that SET of male and female instructors are independent random samples from normally distributed populations with equal variances and possibly different means. As a result, the p-values reported in those studies are for unrealistic null hypotheses and might be misleading.

In contrast, we use permutation tests based on the as-if-random (French natural experiment) or truly random (US experiment) assignment of students to class sections, with no counterfactual assumption that the students, SET scores, grades, or any other variables comprise random samples from any populations, much less populations with normal distributions.

In most cases, our tests are *stratified.* For the US data, for instance, the randomization is stratified on the actual TA: students are randomized within the two sections taught by each TA, but students assigned to different TAs comprise different strata. The randomization is independent across strata. For the French data, the randomization is stratified on course and year: students in different courses or in different years comprise different strata, and the randomization is independent across strata. The null distributions of the test statistics[3] are induced by this random assignment, with no assumption about the distribution of SET or other variables, no parameter

[2] One birth year is obviously incorrect, but our analyses do not rely on the birth years.

[3] The test statistics are correlations of a response variable with experimental variables, or differences in the means of a response variable across experimental conditions, aggregated across strata.

estimates, and no model.

3.1 Illustration: French natural Experiment

The selection of course sections by students at the French university—and the implicit assignment of instructors to sets of students—is as if at random within sections of each course each year, independent across courses and across years. The university's triad system groups students in their classes across disciplines, building small cohorts for each semester. Hence, the randomization for our test keeps these groups of students intact. Stratifying on course topic and year keeps students who took the same final exam grouped in the randomization and honors the design of the natural experiment.

Teaching effectiveness is multidimensional [Marsh and Roche, 1997] and difficult to define, much less measure. But whatever it is, effective teaching should promote student learning: *ceteris paribus*, students of an effective instructor should have better learning outcomes than students of an ineffective instructor have. In the French university, in all courses other than Political Institutions,[4] students in every section of a course in a given year take the same anonymously graded final exam. To the extent that final exams are designed well, scores on these exams reflect relevant learning outcomes for the course. Hence, in each course each semester, students of more effective instructors should do better on the final, on average, than students of less effective instructors.

Consider testing the hypothesis that SET are unrelated to performance on the final exam against the alternative that, all else equal, students of instructors who get higher average SET get higher final exam scores, indicating that they learned more.

[4] The final exam in Political Institutions is oral and hence not graded anonymously.

For this hypothesis test, we omit Political Institutions because the final exam was not anonymous.

The test statistic is the average over courses and years of the Pearson correlation between mean SET and mean final exam score among sections of each course each year. If SET do measure instructors' contributions to learning, we would expect this average correlation to be positive: sections with above-average mean SET in each discipline each year would tend to be sections with above-average mean final exam scores. How surprising is the observed average correlation, if there is no overall connection between mean SET and mean final exam for sections of a course?

There are 950 "individuals," course sections of subjects other than Political Institutions. Each of the 950 course sections has an average SET and an average final exam score. These fall in $3 \times 5 + 2 \times 3 = 21$ year-by-course strata. Under the randomization, within each stratum, instructors are assigned sections independently across years and courses, with the number of sections of each course that each instructor teaches each year held fixed. For instance, if in 2008 there were N sections of History taught by K instructors in all, with instructor k teaching N_k sections, then in the randomization, all

$$\binom{N}{N_1 \cdots N_K} \tag{1}$$

ways of assigning N_k of the N 2008 History sections to instructor k, for $k = 1, \ldots, K$, would be equally likely. The same would hold for sections of other courses and other years. Each combination of assignments across courses and years is equally likely: the assignments are independent across strata.

Under the null hypothesis that SET have no relationship to final exam scores, average final exam scores for sections in each course each year are *exchangeable* given the average SET for the sections. Imagine "shuffling" (i.e., permuting) the average final exam scores across sections of each course each year, independently for different courses and different years. For each permutation, compute the Pearson correlation between average SET for each section and average final exam score for each section, for each course, for each year. Average the resulting 21 Pearson correlations. The probability distribution of that average is the null distribution of the test statistic. The p-value is the upper tail probability beyond the observed value of the test statistic, for that null distribution.

The hypothetical randomization holds triads fixed, to allow for cohort effects and to match the natural experiment. Hence, the test is conditional on which students happen to sign up for which triad. However, if we test at level no greater than α conditionally on the grouping of students into triads, the unconditional level of the resulting test across all possible groupings is no greater than α:

$$\begin{aligned}
\Pr\{\text{ Type I error }\} &= \sum_{\text{all possible sets of triads}} \Pr\{\text{ Type I error } \mid \text{ triads }\} \Pr\{\text{ triads }\} \\
&\leq \sum_{\text{all possible sets of triads}} \alpha \Pr\{\text{triads }\} \\
&= \alpha \sum_{\text{all possible sets of triads}} \Pr\{\text{triads }\} \\
&= \alpha. \qquad (2)
\end{aligned}$$

It is not practical to enumerate all possible permutations of sections within courses and years, so we estimate the p-value by performing 10^5 random permutations within each stratum, finding the value of the test statistic for each overall assignment, and comparing the observed value of the test statistic to the empirical distribution of those 10^5 random values. The probability distribution of the number of random permutations assignments for which the test statistic is greater than or equal to its observed value is Binomial, with n equal to the number of overall random permutations and p equal to the true p-value. Hence, the standard error of the estimated p-values is hence no larger than $(1/2)/\sqrt{10^5} \approx 0.0016$. Code for all our analyses is at `https://github.com/kellieotto/SET-and-Gender-Bias`. Results for the French data are below in section 4.

3.2 Illustration: US Experiment

To test whether perceived instructor gender affects SET in the US experiment, we use the Neyman "potential outcomes" framework [Neyman et al., 1990]. A fixed number N of individuals—e.g., students or classes—are assigned randomly (or as if at random by Nature) into $k \geq 2$ groups of sizes $N_1, \ldots, N_k$. Each group receives a different treatment. "Treatment" is notional. For instance, the treatment might be the gender of the class instructor.

For each individual i, we observe a numerical response R_i. If individual i is assigned to treatment j, then $R_i = r_{ij}$. The numbers $\{r_{ij}\}$ are considered to have been fixed before the experiment. (They are not assumed to be a random sample from any population; they are not assumed to be realizations of any underlying random variables.) Implicit in this notation is the *non-interference* assumption that each individual's response depends only on the treatment that individual receives, and not on which treatments other individuals receive.

We observe only one potential outcome for individual i, depending on which treatment she or he receives. In this model, the responses $\{R_i\}_{i=1}^N$ are random, but only because individuals are assigned to treatments at random, and the assignment determines which of the fixed values $\{r_{ij}\}$ are observed.

In the experiment conducted by MacNell et al. [2014], N students were assigned at random to six sections of an online course, of which four were taught by TAs. Our analysis focuses on the four sections taught by TAs. We condition on the assignment of students to the two sections taught by the professor. Each remaining student i could be assigned to any of $k = 4$ treatment conditions: either of two TAs, each identified as either male or female. The assignment of students to sections was random: each of the

$$\binom{N}{N_1 N_2 N_3 N_4} = \frac{N!}{N_1! N_2! N_3! N_4!} \tag{3}$$

possible assignments of N_1 students to TA 1 identified as male, N_2 student to TA 1 identified as female, etc., was equally likely.

Let r_{i1} and r_{i2} be the ratings student i would give TA 1 when TA 1 is identified as male and as female, respectively; and let r_{i3} and r_{i4} the ratings student i would give TA 2 when that TA is identified as male and as female, respectively. Typically, the null hypotheses we test assert that for each i, some subset of $\{r_{ij}\}$ are equal. For assessing whether the identified gender of the TA affects SET, the null hypothesis is that for each i, $r_{i1} = r_{i2}$ (the rating the ith student would give TA 1 is the same, whether TA 1 is identified as male or female), and $r_{i3} = r_{i4}$ (the rating the ith student would give TA 2 is the same, whether TA 2 is identified as male or female). Different students might give different ratings under the same treatment condition (the null does not assert that $r_{ij} = r_{\ell j}$ for $i \neq \ell$), and the ith student might give different ratings to TA 1 and TA 2 (the null does not assert that $r_{i1} = r_{i3}$). The null hypothesis makes no assertion about the population distributions of $\{r_{i1}\}$ and $\{r_{i3}\}$, nor does it assert that $\{r_{ij}\}$ are a sample from some super-population.

For student i, we observe exactly one of $\{r_{i1}, r_{i2}, r_{i3}, r_{i4}\}$. If we observe r_{i1}, then—if the null hypothesis is true—we also know what r_{i2} is, and vice versa, but we do not know anything about r_{i3} or r_{i4}. Similarly, if we observe either r_{i3} or r_{i4} and the null hypothesis is true, we know the value of both, but we do not know anything about r_{i1} or r_{i2}.

Consider the average SET (for any particular item) given by the N_2+N_4 students assigned to sections taught by an apparently female TA, minus the average SET given by the N_1+N_3 students assigned to sections taught by an apparently male TA. This is what MacNell et al. [2014] tabulate as their key result. If the perceived gender of the TA made no difference in how students rated the TA, we would expect the difference of averages to be close to zero.[5] How "surprising" is the observed difference in averages?

Consider the

$$\binom{N_1+N_2}{N_1} \times \binom{N_3+N_4}{N_3} \tag{4}$$

assignments that keep the same N_1+N_2 students in TA 1's sections (but might change which of those sections a student is in) and the same N_3+N_4 students in TA 2's sections. For each of those assignments, we know what $\{R_i\}_{i=1}^N$ would have been if the null hypothesis is true: each would be exactly the same as its observed value, since those assignments keep students in sections taught by the same TA. Hence, we can calculate the value that the test statistic would have had for each of those assignments.

Because all $\binom{N}{N_1 N_2 N_3 N_4}$ possible assignments of students to sections are equally likely, these $\binom{N_1+N_2}{N_1} \times \binom{N_3+N_4}{N_3}$ assignments in particular are also equally likely. The fraction of those assignments for which the value of the test statistic is at least as large (in absolute value) as the observed value of the test statistic is the p-value of the null hypothesis that students give the same rating (or none) to an TA, regardless

[5] We would expect it to be a least a little different from zero both because of the luck of the draw in assigning students to sections and because students might rate the two TAs differently, regardless of the TA's perceived gender, and the groups are not all the same size.

of the gender that TA appears to have.

This test is conditional on which of the students are assigned to each of the two TAs, but if we test at level no greater than α conditionally on the assignment, the unconditional level of the resulting test across all assignments is no greater than α, as shown above.

In principle, one could enumerate all the equally likely assignments and compute the value of the test statistic for each, to determine the (conditional) null distribution of the test statistic. In practice, there are prohibitively many assignments (for instance, there are $\binom{23}{11}\binom{24}{11} > 3.3 \times 10^{12}$ possible assignments of 47 students to the 4 TA-led sections that keep constant which students are assigned to each TA). Hence, we estimate p-values by simulation, drawing 10^5 equally likely assignments at random, with one exception, noted below. The distribution of the number of simulated assignments for which the test statistic is greater than or equal to its observed value is Binomial with n equal to the number of simulated assignments and p equal to the true p-value. Hence, the standard error of the estimated p-values is hence no larger than $(1/2)/\sqrt{10^5} \approx 0.0016$. Code for all our analyses is at https://github.com/kellieotto/SET-and-Gender-Bias. Results for the US data are in section 5.

4 The French Natural Experiment

In this section, we test hypotheses about relationships among SET, teaching effectiveness, grade expectations, and student and instructor gender. Our tests aggregate data within course sections, to match how SET are typically used in personnel decisions. We use the average of Pearson correlations across strata as the test statistic,[6] which allows us to test both for differences in means (which can be written as correlations with a dummy variable) and for association with ordinal or quantitative variables.

In these analyses, individual i is a section of a course; the "treatment" is the instructor's gender, the average interim grade, or the average final exam score; and the "response" is the average SET or the average final exam score. Strata consist of all sections of a single course in a single year.

Our tests for overall effects stratify on the course subject, to account for systematic differences across departments: the hypothetical randomization shuffles characteristics among courses in a given department, but not across departments. We also perform tests separately in different departments, and in some cases separately by student gender.

4.1 SET and final exam scores

We test whether average SET scores and average final exam scores for course sections are associated (Table 2). The null hypothesis is that the pairing of average final grade and average SET for sections of a course each year is as if at random, independent across courses and across years. We test this hypothesis overall and separately by discipline, using the average Pearson correlation across strata, as described in section 3.1. If the null hypothesis were true, we would expect the test statistic to be close to zero. On the other hand, if SET do measure teaching effectiveness, we would expect average SET and average final exam score to be positively correlated

[6] As discussed above, we find p-values from the (nonparametric) permutation distribution, not from the theoretical distribution of the Pearson correlation under the parametric assumption of bivariate normality.

within courses within years, making the test statistic positive.

The numbers show that SET scores do not measure teaching effectiveness well, overall: the one-sided p-value for the hypothesis that the correlation is zero is 0.09. Separate tests by discipline find that for History, the association is positive and statistically significant (p-value of 0.01), while the other disciplines (Macroeconomics, Microeconomics, Political science and Sociology), the association is either negative or positive but not statistically significant (p-values 0.19, 0.55, 0.62, and 0.61 respectively).

Table 2: Average correlation between SET and final exam score, by subject

	strata	$\bar{\rho}$	p-value
Overall	26 (21)	0.04	0.09
History	5	0.16	0.01
Political Institutions	5	N/A	N/A
Macroeconomics	5	0.06	0.19
Microeconomics	5	-0.01	0.55
Political science	3	-0.03	0.62
Sociology	3	-0.02	0.61

Note: p-values are one-sided, since, if SET measured teaching effectiveness, mean SET should be positively associated with mean final exam scores. Correlations are computed for course-level averages of SET and final exam score within strata, then averaged across strata. Political Institutions is not reported, because the final exam was not graded anonymously. The five strata of Political Institutions are not included in the overall average, which is computed from the remaining 21 strata-level correlation coefficients.

4.2 SET and Instructor Gender

The second null hypothesis we test is that the pairing (by section) of instructor gender and SET is as if at random within courses each year, independently across years and courses. If gender does not affect SET, we would expect the correlation between average SET and instructor gender to be small in each course in each year. On the other hand, if students tend to rate instructors of one gender higher, we would expect the average correlation to be large in absolute value. We find that average SET are significantly associated with instructor gender, with male instructors getting higher ratings (overall p-value 0.00). Male instructors get higher SET on average in every discipline (Table 3) with two-sided p-values ranging from 0.08 for History to 0.63 for Political Science.

Table 3: Average correlation between SET and instructor gender

	$\bar{\rho}$	p-value
Overall	0.09	0.00
History	0.11	0.08
Political institutions	0.11	0.10
Macroeconomics	0.10	0.16
Microeconomics	0.09	0.16
Political science	0.04	0.63
Sociology	0.08	0.34

Note: p-values are two-sided.

4.3 Instructor Gender and Learning Outcomes

Do men receive higher SET scores overall because they are better instructors? The third null hypothesis we test is that the pairing (by course) of instructor gender and average final exam score is as if at random within courses each year, independent across courses and across years. If this hypothesis is true, we would expect the average correlations to be small. If the effectiveness of instructors differs systematically by gender, we would expect average correlation to be large in absolute value. Table 4 shows that on the whole, students of male instructors perform worse on the final than students of female instructors, by an amount that is statistically significant (p-value 0.07 overall). In all disciplines, students of male instructors perform worse, but by amounts that are not statistically significant (p-values ranging from 0.22 for History to 0.70 for Political Science). This suggests that male instructors are not noticeably more effective than female instructors, and perhaps are less effective: The statistically significant difference in SET scores for male and female instructors does not seem to reflect a difference in their teaching effectiveness.

Table 4: Average correlation between final exam scores and instructor gender

	$\bar{\rho}$	p-value
Overall	-0.06	0.07
History	-0.08	0.22
Macroeconomics	-0.06	0.37
Microeconomics	-0.06	0.37
Political science	-0.03	0.70
Sociology	-0.05	0.55

Note: p-values are two-sided. Negative values of $\bar{\rho}$ indicate that students of female instructors did better on average than students of male instructors.

4.4 Gender Interactions

Why do male instructors receive higher SET scores? Separate analyses by student gender shows that male students tend to give higher SET scores to male instructors (Table 5). These permutation tests confirm the results found by Boring [2015a]. Gender concordance is a good predictor of SET scores for men (p-value 0.00 overall). Male students give significantly higher SET scores to male instructors in History (p-value 0.01), Microeconomics (p-value 0.01), Macroeconomics (p-value 0.04), Political Science (p-value 0.06), and Political Institutions (p-value 0.08). Male students give higher SET scores to male instructors in Sociology as well, but the effect is not statistically significant (p-value 0.16).

The correlation between gender concordance and overall satisfaction scores for female students is also positive overall and weakly significant (p-value 0.09). The correlation is negative in some fields (History, Political Institutions, Macroeconomics, Microeconomics and Sociology) and positive in only one field (Political Science), but in no case statistically significant (p-values range from 0.12 to 0.97).

Table 5: Average correlation between SET and gender concordance

	Male student		Female student	
	$\bar{\rho}$	p-value	$\bar{\rho}$	p-value
Overall	0.15	0.00	0.05	0.09
History	0.17	0.01	-0.03	0.60
Political institutions	0.12	0.08	-0.11	0.12
Macroeconomics	0.14	0.04	-0.05	0.49
Microeconomics	0.18	0.01	-0.00	0.97
Political science	0.17	0.06	0.04	0.64
Sociology	0.12	0.16	-0.03	0.76

Note: p-values are two-sided.

Do male instructors receive higher SET scores from male students because their teaching styles match male students' learning styles? If so, we would expect male students of male instructors to perform better on the final exam. However, they do not (Table 6). If anything, male students of male instructors perform worse overall on the final exam (the correlation is negative but not statistically significant, with a p-value 0.75). In History, the amount by which male students of male instructors do worse on the final is significant (p-value 0.03): male History students give significantly higher SET scores to male instructors, despite the fact that they seem to learn more from female instructors. SET do not appear to measure teaching effectiveness, at least not primarily.

4.5 SET and grade expectations

The next null hypothesis we test is that the pairing by course of average SET scores with average interim grades is as if at random. Because interim grades may set student grade expectations, if students give higher SET in courses where they expect

Table 6: Average correlation between student performance and gender concordance

	Male student		Female student	
	$\bar{\rho}$	p-value	$\bar{\rho}$	p-value
Overall	-0.01	0.75	0.06	0.07
History	-0.15	0.03	-0.02	0.74
Macroeconomics	0.04	0.60	0.11	0.10
Microeconomics	0.02	0.80	0.07	0.29
Political science	0.08	0.37	0.11	0.23
Sociology	0.01	0.94	0.06	0.47

Note: p-values are two-sided.

higher grades, the association should be positive. Indeed, the association is positive and generally highly statistically significant (Table 7). Political institutions is the only discipline for which the average correlation between interim grades and SET scores is negative, but the correlation is not significant (p-value 0.61). The estimated p-values for all other courses are between 0.0 and 0.03. The average correlations are especially high in History (0.32) and Sociology (0.24).

Table 7: Average correlation between SET and interim grades

	$\bar{\rho}$	p-value
Overall	0.16	0.00
History	0.32	0.00
Political institutions	-0.02	0.61
Macroeconomics	0.15	0.01
Microeconomics	0.13	0.03
Political science	0.17	0.02
Sociology	0.24	0.00

Note: p-values are one-sided.

In summary, the average correlation between SET and final exam grades (at the level of class sections) is positive, but only weakly significant overall and not significant for most disciplines. However, the average correlation between SET and grade expectations (at the level of class sections) is positive and significant overall and across most disciplines. The average correlation between instructor gender and SET is statistically significant—male instructors get higher SET—but if anything, students of male instructors do worse on final exams than students of female instructors. Male students tend to give male instructors higher SET, even though they might be learning less than they do from female instructors. We conclude that SET are influenced more by instructor gender and student grade expectations than by teaching effectiveness.

5 The US Randomized Experiment

The previous section suggests that SET have little connection to teaching effectiveness, but the natural experiment does not allow us to control for differences in teaching styles across instructors. MacNell et al. [2014] does. As discussed above, MacNell et al. [2014] collected SET from an online course in which 43 students were randomly assigned to four[7] discussion groups, each taught by one of two TAs, one male and one female. The TAs gave similar feedback to students, returned assignments at exactly the same time, etc.

Biases in student ratings are revealed by differences in ratings each TA received when that TA is identified to the students as male versus as female. MacNell et al. [2014] find that "the male identity received significantly higher scores on professionalism, promptness, fairness, respectfulness, enthusiasm, giving praise, and the student ratings index . . . Students in the two groups that perceived their assistant instructor to be male rated their instructor significantly higher than did the students in the two groups that perceived their assistant instructor to be female, regardless of the actual

[7] As discussed above, there were six sections in all, of which two were taught by the professor and four were taught by TAs.

gender of the assistant instructor." MacNell et al. [2014] used parametric tests whose assumptions did not match their experimental design; part of our contribution is to show that their data admit a more rigorous analysis using permutation tests that honor the underlying randomization and that avoid parametric assumptions about SET. The new analysis supports their overall conclusions, in some cases substantially more strongly than the original analysis (for instance, p-values of 0.01 versus 0.19 for promptness and fairness). In other cases, the original parametric tests overstated the evidence (for instance, a p-value of 0.29 versus 0.04 for knowledgeability).

We use permutation tests as described above in section 3. Individual i is a student; the treatment is the combination of the TA's identity and the TA's apparent gender (there are $K = 4$ treatments). The null hypothesis is that each student would give a TA the same SET score, whether that TA is apparently male or apparently female. A student might give the two TAs different scores, and different students might give different scores to the same TA.

Because of how the experimental randomization was performed, all allocations of students to TA sections that preserve the number of students in each section are equally likely, including allocations that keep the same students assigned to each actual TA constant.

To test whether there is a systematic difference in how students rate apparently male and apparently female TAs, we use the difference in pooled means as our test statistic: We pool the SET for both instructors when they are identified as female and take the mean, pool the SET for both instructors when they are identified as male and take the mean, then subtract the second mean from the first mean (Table 8). This is what MacNell et al. [2014] report as their main result.

As described above, the randomization is stratified and conditions on the set of students allocated to each TA, because, under the null hypothesis, we then know what

SET students would have given for each possible allocation, completely specifying the null distribution of the test statistic. The randomization includes the nonresponders, who are omitted from the averages of the group they are assigned to.

We also perform tests involving the association of concordance of student and apparent TA gender, (Table 9) and SET and concordance of student and actual TA gender (Table 10) using the pooled difference in means as the test statistic. We test the association between grades and actual TA gender (Table 11) using the average Pearson correlation across strata as the test statistic. We find the p-values from the stratified permutation distribution of the test statistic, avoiding parametric assumptions.

5.1 SET and Perceived Instructor Gender

The first hypothesis we test is that students would rate a given TA the same, whether the student thinks the TA is female or male. A positive value of the test statistic means that students give higher SET on average to apparently male instructors. There is weak evidence that the overall SET score depends on the perceived gender (p-value 0.12). The evidence is stronger for several other items students rated: fairness (p-value 0.01), promptness (p-value 0.01), giving praise (p-value 0.01), enthusiasm (p-value 0.06), communication (p-value 0.07), professionalism (p-value 0.07), respect (p-value 0.06), and caring (p-value 0.10). For seven items, the nonparametric permutation p-values are smaller than the parametric p-values reported by MacNell et al. [2014]. Items for which the permutation p-values were greater than 0.10 include clarity, consistency, feedback, helpfulness, responsiveness, and knowledgeability. SET were on a 5-point scale, so a difference in means of 0.80, observed in student ratings of the promptness with which assignments were returned, is 16%

of the full scale—an enormous difference. Since assignments were returned at exactly the same time in all four sections of the class, this seriously impugns the ability of SET to measure even putatively objective characteristics of teaching.

Table 8: Mean ratings and reported instructor gender (male minus female)

	difference in means	nonparametric *p*-value	MacNell et al. *p*-value
Overall	0.47	0.12	0.128
Professional	0.61	0.07	0.124
Respectful	0.61	0.06	0.124
Caring	0.52	0.10	0.071
Enthusiastic	0.57	0.06	0.112
Communicate	0.57	0.07	NA
Helpful	0.46	0.17	0.049
Feedback	0.47	0.16	0.054
Prompt	0.80	0.01	0.191
Consistent	0.46	0.21	0.045
Fair	0.76	0.01	0.188
Responsive	0.22	0.48	0.013
Praise	0.67	0.01	0.153
Knowledge	0.35	0.29	0.038
Clear	0.41	0.29	NA

Note: p-values are two-sided.

We also conducted separate tests by student gender. In contrast to our findings for the French data, where male students rated male instructors higher, in the MacNell et al. [2014] experiment, perceived male instructors received significantly higher evaluation scores because female students rated the perceived male instructors higher (Table 9). Male students rated the perceived male instructor significantly (though weakly) higher on only one criterion: fairness (*p*-value 0.09). Female students, however, rated the perceived male instructor higher on overall satisfaction (*p*-value 0.11) and most teaching dimensions: praise (*p*-value 0.01), enthusiasm (*p*-value 0.05), caring (*p*-value 0.05), fairness (*p*-value 0.04), respectfulness (*p*-value 0.12), communication (*p*-value 0.10), professionalism (*p*-value 0.12), and feedback (*p*-value 0.10).

Female students rate (perceived) female instructors lower on helpfulness, promptness, consistency, responsiveness, knowledge, and clarity, although the differences are not statistically significant.

Table 9: SET and reported instructor gender (male minus female)

	Male students		Female students	
	difference in means	*p*-value	difference in means	*p*-value
Overall	0.17	0.82	0.79	0.11
Professional	0.42	0.55	0.82	0.12
Respectful	0.42	0.55	0.82	0.12
Caring	0.04	1.00	0.96	0.05
Enthusiastic	0.17	0.83	0.96	0.05
Communicate	0.25	0.68	0.87	0.10
Helpful	0.46	0.43	0.51	0.35
Feedback	0.08	1.00	0.88	0.10
Prompt	0.71	0.15	0.86	0.13
Consistent	0.17	0.85	0.77	0.17
Fair	0.75	0.09	0.88	0.04
Responsive	0.38	0.54	0.06	1.00
Praise	0.58	0.29	0.81	0.01
Knowledge	0.17	0.84	0.54	0.21
Clear	0.13	0.85	0.67	0.29

Note: p-values are two-sided.

Students of both genders rated the apparently male instructor higher on all dimensions, by an amount that often was statistically significant for female students (Table 9). However, students rated the actual male instructor higher on some dimensions and lower on others, by amounts that generally were not statistically significant (Table 10). The exceptions were praise (*p*-value 0.02) and responsiveness (*p*-value 0.05), where female students tended to rate the actual female instructor significantly higher.

Students of the actual male instructor performed worse in the course on average, by an amount that was statistically significant (Table 11). The difference in student performance by perceived gender of the instructor is not statistically significant.

Table 10: SET and actual instructor gender (male minus female)

	Male students		Female students	
	difference in means	*p*-value	difference in means	*p*-value
Overall	-0.13	0.61	-0.29	0.48
Professional	0.15	0.96	-0.09	0.73
Respectful	0.15	0.96	-0.09	0.73
Caring	-0.22	0.52	-0.07	0.75
Enthusiastic	-0.13	0.62	-0.44	0.29
Communicate	-0.02	0.80	-0.18	0.61
Helpful	0.03	0.89	0.26	0.71
Feedback	-0.24	0.48	-0.41	0.36
Prompt	-0.09	0.69	-0.33	0.44
Consistent	0.12	0.97	-0.40	0.35
Fair	-0.06	0.71	-0.59	0.12
Responsive	-0.13	0.64	-0.68	0.05
Praise	0.02	0.86	-0.60	0.02
Knowledge	0.22	0.83	-0.44	0.17
Clear	-0.26	0.49	-0.98	0.07

Note: p-values are two-sided.

Table 11: Mean grade and instructor gender (male minus female)

	difference in means	*p*-value
Perceived	1.76	0.54
Actual	-6.81	0.02

Note: p-values are two-sided.

These results suggest that students rate instructors more on the basis of the instructor's perceived gender than on the basis of the instructor's effectiveness. Students of the TA who is actually female did substantially better in the course, but students rated apparently male TAs higher.

6 Multiplicity

We did not adjust the p-values reported above for multiplicity. We performed a total of approximately 50 tests on the French data, of which we consider four to be our primary results:

1FR lack of association between SET and final exam scores (a negative result, so multiplicity is not an issue)

2FR lack of association between instructor gender and final exam scores (a negative result, so multiplicity is not an issue)

3FR association between SET and instructor gender

4FR association between SET and interim grades

Bonferroni's adjustment for these four tests would leave the last two associations highly significant, with adjusted p-values less than 0.01.

We performed a total of 77 tests on the US data. We consider the three primary null hypotheses to be

1US perceived instructor gender plays no role in SET

2US male students rate perceived male and female instructors the same

3US female students rate perceived male and female instructors the same

To account for multiplicity, we tested these three "omnibus" hypotheses using the nonparametric combination of tests (NPC) method with Fisher's combining function [Pesarin and Salmaso, 2010, Chapter 4] to summarize the 15 dimensions of teaching into a single test statistic that measures how "surprising" the 15 observed differences would be for each of the three null hypotheses. In 10^5 replications, the empirical p-values for these three omnibus hypotheses were 0 (99% confidence interval $[0.0, 5.3 \times 10^{-5}]$), 0.464 (99% confidence interval $[0.460, 0.468]$), and 0 (99% confidence interval $[0.0, 5.3 \times 10^{-5}]$), respectively. (The confidence bounds were obtained by inverting Binomial hypothesis tests.) Thus, we reject hypotheses 1US and 3US.

We made no attempt to optimize the tests to have power against the alternatives considered. For instance, with the US data, the test statistic grouped the two identified-as-female sections and the two identified-as-male conditions, in keeping with how MacNell et al. [2014] tabulated their results, rather than using each TA as his or her own control (although the randomization keeps the two strata intact). Given the relatively small number of students in the US experiment, it is remarkable that *any* of the p-values is small, much less that the p-values for the omnibus tests are effectively zero.

7 Code and Data

Jupyter (`http://jupyter.org/`) notebooks containing our analyses are at `https://github.com/kellieotto/SET-and-Gender-Bias`; they rely on the `permute` Python library (`https://pypi.python.org/pypi/permute/`). The US data are available at `http://n2t.net/ark:/b6078/d1mw2k`. French privacy law prohibits publishing the French data.

8 Discussion

8.1 Other studies

To our knowledge, only two experiments have controlled for teaching style in their designs: Arbuckle and Williams [2003] and MacNell et al. [2014]. In both experiments, students generally gave higher SET when they *thought* the instructor was male, regardless of the actual gender of the instructor. Both experiments found that systematic differences in SET by instructor gender reflect gender bias rather than a match of teaching style and student learning style or a difference in actual teaching effectiveness.

Arbuckle and Williams [2003] showed a group of 352 students "slides of an age- and gender-neutral stick figure and listened to a neutral voice presenting a lecture and then evaluated it on teacher evaluation forms that indicated 1 of 4 different age and gender conditions (male, female, 'old,' and 'young')" [Arbuckle and Williams, 2003, p.507]. All students saw the same stick figure and heard the same voice, so differences in SET could be attributed to the age and gender the students were *told* the instructor had. When students were told the instructor was young and male, students rated the instructor higher than for the other three combinations, especially on "enthusiasm," "showed interest in subject," and "using a meaningful voice tone."

Instructor race is also associated with SET. In the US, SET of instructors of color appear to be biased downwards: minority instructors tend to receive significantly lower SET scores compared to white (male) instructors [Merritt, 2008].[8] Age, [Arbuckle and Williams, 2003], charisma [Shevlin et al., 2000], and physical attractiveness [Riniolo et al., 2006, Hamermesh and Parker, 2005] are also associated

[8] French law does not allow the use of race-related variables in data sets. We were thus unable to test for racial biases in SET using the French data.

with SET. Other factors generally not in the instructor's control that may affect SET scores include class time, class size, mathematical or technical content, and the physical classroom environment [Hill and Epps, 2010].

Many studies cast doubt on the validity of SET as a measure of teaching effectiveness (see Johnson [2003, Chapters 3–5] for a review and analysis, Pounder [2007] for a review, and Galbraith et al. [2012], Carrell and West [2010] for exemplars). Some studies find that gender and SET are not significantly associated [Bennett, 1982, Centra and Gaubatz, 2000, Elmore and LaPointe, 1974] and that SET are valid and reliable measures of teaching effectiveness [Benton and Cashin, 2012, Centra, 1977].[9] The contradictions among conclusions suggests that if SET are ever valid, they are not valid in general: universities should not assume that SET are broadly valid at their institution, valid in any particular department, or valid for any particular course. Given the many sources of bias in SET and the variability in magnitude of the bias by topic, item, student gender, etc., as a practical matter it is impossible to adjust for biases to make SET a valid, useful measure of teaching effectiveness.

8.2 Summary

We used permutation tests to examine data collected by Boring [2015a] and MacNell et al. [2014], both of which find that gender biases prevent SET from measuring teaching effectiveness accurately and fairly. SET are more strongly related to instructor's perceived gender and to students' grade expectations than they are to learning, as measured by performance on anonymously graded, uniform final exams. The extent and direction of gender biases depend on context, so it is impossible to adjust for such biases to level the playing field. While the French university data show a

[9] Some authors who claim that SET are valid have a financial interest in developing SET instruments and conducting SET.

positive male student bias for male instructors, the experimental US setting suggests a positive female student bias for male instructors. The biases in the French university data vary by course topic; the biases in the US data vary by item. We would also expect the bias to depend on class size, format, level, physical characteristics of the classroom, instructor ethnicity and a host of other variables.

We do not claim that there is *no* connection between SET and student performance. However, the observed association is sometimes positive and sometimes negative, and in general is not statistically significant—in contrast to the statistically significant strong associations between SET and grade expectations and between SET and instructor gender. SET appear to measure student satisfaction and grade expectations more than they measure teaching effectiveness [Stark and Freishtat, 2014, Johnson, 2003]. While student satisfaction may *contribute* to teaching effectiveness, it is not itself teaching effectiveness. Students may be satisfied or dissatisfied with courses for reasons unrelated to learning outcomes—and not in the instructor's control (e.g., the instructor's gender).

In the US, SET have two primary uses: instructional improvement and personnel decisions, including hiring, firing, and promoting instructors. We recommend caution in the first use, and discontinuing the second use, given the strong student biases that influence SET, even on "objective" items such as how promptly instructors return assignments.[10]

[10] In 2009, the French Ministry of Higher Education and Research upheld a 1997 decision of the French State Council that public universities can use SET only to help tenured instructors improve their pedagogy, and that the administration may not use SET in decisions that might affect tenured instructors' careers (c.f. Boring [2015b]).

8.3 Conclusion

In two very different universities and in a broad range of course topics, SET measure students' gender biases better than they measure the instructor's teaching effectiveness. Overall, SET disadvantage female instructors. There is no evidence that this is the exception rather than the rule. Hence, the onus should be on universities that rely on SET for employment decisions to provide convincing affirmative evidence that such reliance does not have disparate impact on women, under-represented minorities, or other protected groups. Because the bias varies by course and institution, affirmative evidence needs to be specific to a given course in a given department in a given university. Absent such specific evidence, SET should not be used for personnel decisions.

References

J. Arbuckle and B. D. Williams. Students' Perceptions of Expressiveness : Age and Gender Effects on Teacher Evaluations. *Sex Roles*, 49(November):507–516, 2003.

S. K. Bennett. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, 74(2):170–179, 1982.

S. L. Benton and W. E. Cashin. Student ratings of teaching: A summary of research and literature. IDEA Paper 50, The IDEA Center, 2012.

A. Boring. Gender biases in student evaluations of teachers. Document de travail OFCE 13, OFCE, April 2015a.

A. Boring. Can students evaluate teaching quality objectively? Le blog de l'ofce, OFCE, 2015b. URL http://www.ofce.sciences-po.fr/blog/can-students-evaluate-teaching-quality-objectively/.

M. Braga, M. Paccagnella, and M. Pellizzari. Evaluating students evaluations of professors. *Economics of Education Review*, 41:71–88, 2014.

S. E. Carrell and J. E. West. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, 118(3):409–432, June 2010. ISSN 0022-3808. doi: 10.1086/653808. URL http://www.jstor.org/stable/10.1086/653808.

J. A. Centra. Student ratings of instruction and their relationship to student learning. *American educational research journal*, 14(1):17–24, 1977.

J. A. Centra and N. B. Gaubatz. Is There Gender Bias in Student Evaluations of Teaching? *Journal of Higher Education*, 71(1):17–33, 2000. URL http://www.jstor.org/stable/10.2307/2649280.

P. B. Elmore and K. A. LaPointe. Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology*, 66(3):386–389, 1974.

C. S. Galbraith, G. B. Merrill, and D. M. Kline. Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? a neural network and bayesian analyses. *Research in Higher Education*, 53(3):353–374, 2012.

D. S. Hamermesh and A. Parker. Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity. *Economics of Education Review*, 24(4): 369–376, 2005.

M. C. Hill and K. K. Epps. The impact of physical classroom environment on student satisfaction and student evaluation of teaching in the university environment. *Academy of Educational Leadership Journal*, 14(4):65–79, 2010.

V. E. Johnson. *Grade Inflation: A Crisis in College Education.* Springer-Verlag, New York, 2003.

L. MacNell, A. Driscoll, and A. N. Hunt. Whats in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, pages 1–13, 2014.

H. W. Marsh and L. A. Roche. Making Students' Evaluations of Teaching Effectiveness Effective. *American Psychologist*, 52(11):1187–1197, 1997.

D. J. Merritt. Bias, the brain, and student evaluations of teaching. *St. John's Law Review*, 81(1):235–288, 2008.

J. Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

F. Pesarin and L. Salmaso. *Permutation Tests for Complex Data: Theory, Applications and Software.* Wiley, New York, 2010.

J. S. Pounder. Is student evaluation of teaching worthwhile?: An analytical framework for answering the question. *Quality Assurance in Education*, 15(2):178–191, 2007. ISSN 0968-4883. doi: 10.1108/09684880710748938. URL http://www.emeraldinsight.com/10.1108/09684880710748938.

T. C. Riniolo, K. C. Johnson, T. R. Sherman, and J. A. Misso. Hot or not: do professors perceived as physically attractive receive higher student evaluations? *The Journal of general psychology*, 133(1):19–35, Jan. 2006. ISSN 0022-1309. doi: 10.3200/GENP.133.1.19-35. URL http://www.ncbi.nlm.nih.gov/pubmed/16475667.

M. Shevlin, P. Banyard, M. Davies, and M. Griffiths. The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education*, 25(4):397–405, 2000.

P. B. Stark and R. Freishtat. An evaluation of course evaluations. *Science Open Research*, 2014. doi: 10.14293/S2199-1006.1.-.AOFRQA.v1. URL https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015-dc99b467b6e4.

Quality Assurance in Education

Is student evaluation of teaching worthwhile?: An analytical framework for answering the question

James S. Pounder

Is student evaluation of teaching worthwhile?

An analytical framework for answering the question

James S. Pounder
Higher Colleges of Technology, Abu Dhabi, United Arab Emirates

Abstract

Purpose – To present a framework to facilitate comprehension of research on the effectiveness of the teaching evaluation process.

Design/methodology/approach – A comprehensive review of the literature that identifies common categories and factors that can be used to construct an analytical framework.

Findings – Identifies student related, course related and teacher related aspects of research on teaching evaluations. Factors commonly addressed within these aspects are also identified.

Research limitations/implications – Use of the framework to analyse the literature on the student evaluation of teaching (SET) process leads to the view that the time is right to explore other methods of assessing classroom dynamics that could supplement the conventional teacher evaluation process.

Practical implications – Educational literature is replete with studies of the SET system, yet due to the preponderance of these studies, it is difficult to take an overview on the effectiveness of this system. On the basis of a comprehensive survey of the literature, this paper identifies and discusses the central factors influencing SET scores. These factors are then presented in a comprehensible table that can be used as a reference point for researchers and practitioners wishing to examine the effectiveness of the SET system.

Originality/value – The paper is one of the few to attempt to make sense of the myriad of studies on teacher evaluation and to develop a framework to facilitate analysis of the effectiveness of the SET system.

Keywords Students, Training evaluation, Classrooms, Leadership

Paper type Research paper

1. Introduction

Student evaluation of teaching (SET) is a widely used instrument in higher education. For example, Seldin (1993) noted an 86 per cent use of the student evaluation of teaching (SET) as a central feature of personnel decisions in US higher education, an increase in usage from 68 per cent in 1984 and 28 per cent in 1973 (Seldin, 1984). In a feature for the *Chronicle of Higher Education*, Wilson (1998, p. A12) stated that:

> . . . only about 30 per cent of colleges and universities asked students to evaluate professors in 1973, but it is hard to find an institution that doesn't today. Such evaluations are now the most important, and sometimes the sole, measure of a teacher's teaching ability.

The extent of reliance on the SET as the predominant measure of university teacher performance is not confined to the USA; it is a worldwide phenomenon (Newton, 1988; Seldin, 1989; Stratton, 1990).

Arguably, the heavy reliance on the SET would be justified if ratings of teacher performance were generally reflected in student achievement. However, there is

Quality Assurance in Education
Vol. 15 No. 2, 2007
pp. 178-191
© Emerald Group Publishing Limited
0968-4883
DOI 10.1108/09684880710748938

considerable disagreement in the literature on the link between SET scores and student achievement. Despite the existence of studies indicating that SET's are reasonably valid multidimensional measures (Marsh and Roche, 1997; McKeachie, 1987) and have a moderate correlation with student learning (d'Apollonia and Abrami, 1997), by and large, most investigations have found little correlation between student achievement and student ratings of their teachers. Cohen's (1983) meta-analysis, for example, found that student achievement accounted for only 14.4 per cent of overall teacher rating variance. Similarly, a meta-analysis by McCallum (1984) found that student achievement explained only 10.1 per cent of overall teacher rating variance. Equally, a 1982 investigation by Dowell and Neal revealed that student achievement accounted for only 3.9 per cent of between-teacher student rating variance (Dowell and Neal, 1982). Finally, a comprehensive study by Damron (1996) found that most of the factors contributing to student ratings of university teachers are probably unrelated to a teacher's ability to promote student learning.

It is findings such as those presented above that have led commentators such as Reckers (1995, p. 33) to state that:

> ... nearly 75 per cent of academics judge student course evaluations as unreliable and imprecise metrics of performance, yet nearly 100 per cent of schools use them, frequently exclusively.

The remainder of this paper presents a framework for examining research on the factors influencing SET scores that lends support to the view that the typical SET system is seriously flawed.

2. A framework for analysis

The literature is replete with studies of the SET phenomenon (Wilson, 1998) and analysis of the findings indicates a triad comprising student related factors, course related factors, and teacher related factors. This triad is presented in summary form in Table I and is followed by a description of the various factors within the student related, course related and teacher related categories. Arguably, the literature on teacher evaluation generally falls within one or more of these categories and tends to address one or more of the factors subsumed within these categories. Consequently, Table I presents a useful framework for making sense of the myriad of research studies on the SET system.

2. 1. Student related factors

Studies tend to revolve around student gender in terms of the extent to which male or female students generally give higher or lower SET scores. Additionally, a few studies have examined the effect of student academic level and maturity on SET scoring. Further, one study has suggested that students use the SET to punish teachers who are perceived to be working them too hard or who have given them low grades. Each of these factors is discussed in more detail below.

Gender effect. More than one study has indicated that student ratings of teachers are influenced by student gender. For example, the study of Walumbwa and Ojode (2000), carried out in a US university, indicated that females, particularly at the undergraduate level, rated their classroom teachers generally higher on classroom leadership dimensions than did their male counterparts. Bachen *et al.* (1999) found a strong

Table I.
Factors influencing SET scores

Student related factors		
Gender	Generally higher SET scores by male or female students	Bachen *et al.* (1999); Walumbwa and Ojode (2000); Feldman (1993)
Academic level and maturity	SET score related to academic level of course or student maturity	Aleamoni (1981); Frey *et al.* (1975); Holtfreter (1991); Langbein (1994); Marsh (1984)
Punishing teachers for low grades		Crumbley *et al.* (2001)
Course related factors		
Grading	High SET scores for high grades or high grade expectations	Aronson and Linder (1965); Brown (1976); Centra and Creech (1976); Goldman (1993); Greenwald (1997); Johnson and Christian (1990); Perkins *et al.* (1990)
Class size		Feldman (1984); Glass *et al.* (1981); Holtfreter (1991); Koh and Tan (1997); Liaw and Goh (2003); Langbein (1994); Marsh (1987); Meredith (1984); Toby (1993)
Course content	SET scores influenced by academic discipline, degree of course difficulty, required versus elective	Cashin (1990); Clark (1993); Cranton and Smith (1986); Aleamoni (1989); DeBerg and Wilson (1990); Stodolsky (1984)
Class timing	SET scores influenced by timing of teaching evaluation when timing of evaluation depends on timing of the course	Cronin and Capie (1986); DeBerg and Wilson (1990); Husbands and Fosh (1993); Koh and Tan (1997)
Teacher related factors		
Gender	SET scores influenced by the gender of the teacher	Bennett (1982); Cooper *et al.* (1982); Crawford and MacLeod (1990); Downs and Downs (1993); Feldman (1993); Langbein (1994); Rubin (1981); Sears and Hennessey (1996); Kierstead *et al.* (1988); Winocur *et al.* (1989)
Age, experience and rank (of teacher)		Clayson (1999); Feldman (1983); Holtfreter (1991); Langbein (1994); Smith and Kinney (1992)
Teachers' influencing tactics	Grade inflation, leniency, bringing food to class on the day of the evaluations etc)	Bauer (1996); Crumbley (1995); Crumbley *et al.* (2001); Emery (1995); Handlin (1996); Martin (1998); Powell (1977); Ryan *et al.* (1980); Sacks (1996); Hocutt (1987-1988); Simpson and Siguaw (2000); Stumpf and Freedman (1979); Winsor (1977); Worthington and Wong (1979); Yunker and Marlin (1984)
Teachers' behavioural traits	The "likeability" factor	Abrami *et al.* (1982); Cardy and Dobbins (1986); Feldman (1986); Jackson *et al.* (1999); Naflulin *et al.* (1973); Williams and Ceci (1997)

interaction between student gender and professor gender with female students giving especially high ratings to female professors and comparatively lower ratings to male professors on measures reflecting the qualities of being caring-expressive, interactive, professional-challenging, and organized. By contrast, in the same study, the evaluations by male students of male and female professors did not differ significantly on any of these factors. Bachen *et al.*'s (1999) study confirmed similar findings by Feldman (1993).

Student's academic level and maturity. Frey *et al.* (1975) found that more experienced students were clearly more lenient in their ratings than their younger counterparts. Langbein (1994) suggested that higher level students (i.e. those taking higher level courses) are generally more motivated and discriminating in their evaluation of teaching than lower level students. The implication that SET results will tend to be more favourable for higher level subjects has been confirmed by Marsh (1984) and Holtfreter (1991). Further, Aleamoni's (1981) review of prior research cited eight studies that showed no significant relationship between SET results and student level and 18 studies that reported a positive and significant relationship between these two variables. Furthermore, it is interesting to note that Walumbwa and Ojode's (2000) study, referred to earlier, did reveal differences in sensitivity to classroom leadership qualities between the undergraduate and graduate samples.

Students punishing their teachers via SET scores. It is expected that students will use the SET to reflect back to their teachers and the institutions in question, poor teaching performance. However, Crumbley *et al.* (2001), in their examination of students' perception of the evaluation system, discovered that poor SET scores may reflect as much the inadequacy of student effort as they do the quality of the instruction they have received. Thus, Crumbley *et al.* (2001) found that students will punish their teachers via the SET for being asked embarrassing questions (i.e. questions for which the student has no answer), for being graded hard, for being given quizzes and for being given significant homework. Therefore, the SET can be used as a vehicle for students to punish conscientious educators.

2.2. Course related factors

The central area that has received attention is the relationship between grades expected by, or awarded to students and SET scores. Quite simply, there is a sizeable body of work indicating that SET scores are sensitive to grade levels and in particular expected grade levels. Other course related aspects that continue to interest researchers in terms of their effect on SET scores are class size, the nature of the course (i.e. degree of perceived content difficulty, core or elective course etc), and the timing of course delivery (i.e. end of day/week) insofar as this affects the timing of the evaluations. Details of the research findings on these course related aspects are presented below.

Grading. One of the key course related areas that has been investigated in relation to SET scores is the influence of actual grading and students' expectations of grades on SET's. Perkins *et al.* (1990) concluded there was evidence that SET scores were sensitive to the grades professors assigned although Johnson and Christian (1990) noted that expected grades were more highly correlated than assigned grades with student ratings. Nevertheless, both studies confirmed that students with higher than expected grades gave higher SET scores than those with lower than expected grades. While Brown (1976) found that grades accounted for only 9 per cent of variation in student ratings, he found that grades were substantially more influential than other factors expected to correlate with student ratings. Greenwald (1997) on the other hand, found that grades distort ratings away from the valid measurement of instructional quality by amounts as much as 20 per cent of ratings variance. Centra and Creech (1976) also found a significant correlation between student grade expectations and SET mean rating scores. Therefore, in practice, students are likely to give high ratings in appreciation of high grades (Aronson and Linder, 1965; Goldman, 1993) or the expectation of high grades irrespective of whether these high grades or expectations actually reflect high academic attainment.

Class size. Student ratings of university teachers have been found to vary with class size (Meredith, 1984; Toby, 1993) and, with a few exceptions (e.g. Langbein, 1994; Marsh, 1987), this is one of the most consistent findings in the literature (Koh and Tan, 1997). In general, smaller class sizes tend to result in better SET scores (Feldman, 1984; Holtfreter, 1991; Koh and Tan, 1997; Liaw and Goh, 2003) probably because the opportunity for teacher-student interaction and rapport is greater in smaller sized classes than larger ones (Glass *et al.*, 1981; Toby, 1993). There is, however, a non-linear relationship between class size and SET scores with both relatively small and relatively large classes receiving better ratings (Feldman, 1984; Holtfreter, 1991).

Course content. Stodolsky (1984) has argued that some courses are more difficult to teach than others and thus, course content is likely to influence SET results. Stodolsky's contention is supported by Clark (1993), DeBerg and Wilson (1990) and Cranton and Smith (1986). In contrast, Langbein (1994), despite noting that there is a general perception that teachers delivering "hard" quantitative subjects are likely to receive lower student ratings than those teaching "soft" qualitative subjects, found no evidence of a significant relationship between type of course and overall teaching ratings. However, in a Singaporean setting, Koh and Tan (1997) found that, in a three-year undergraduate business programme, better SET results were associated with first and third year courses than with second year courses. Student academic level and maturity (discussed above) is given as a possible explanation for the third year SET scores and the authors have offered relative ease of learning introductory courses plus student prior familiarity with course content via pre-university studies as likely explanations of the first year phenomenon. They also noted that the nature of the programme under study could have had a significant influence on their results because the programme required students to undertake a particular specialized field in the second year that could prove challenging and that this might account for the relatively lower SET results for courses taken in the second year.

Cashin (1990) examined very large databases of students' ratings and found significant differences in how students rate teaching across various academic disciplines. Hence, arts and humanities courses tend to receive the highest student ratings, biological and social sciences and health and other professions fall into the medium group, English language and literature and history both fall into the medium-low group with business, economics, computer science, mathematics, the physical sciences and engineering falling in the bottom group. Finally, Aleomoni (1989) observed a rating bias against required courses as opposed to elective courses and noted that the more students in a class taking a required course, the lower the relevant SET score, presumably a feature of the interaction of required course and class size (discussed above).

Class timing. Cronin and Capie (1986) found that teaching evaluation results vary from day to day. Thus, to the extent that evaluations are conducted during the classes in question, the timing of classes is a factor affecting SET results. DeBerg and Wilson (1990) and Husbands and Fosh (1993) have suggested that the time and day a course is taught can affect SET results and in a Singaporean university business school context, Koh and Tan (1997) found that SET's conducted in the later part of the week seemed to result in better teaching evaluations. Koh and Tan have speculated that a more relaxed atmosphere exists towards the end of the week that might have a positive effect on SET scores.

2.3. Teacher related factors

A central theme in teacher related research is the effect of teacher gender including the influence of gender role expectations on teaching evaluations. Additionally a teacher related dimension that continues to provide a focus for research is what is termed here and in the framework (Table I), teacher influencing tactics. A particular feature of this, for example, is deliberate grade inflation in order to "court" high SET scores. In discussing course related factors earlier in this paper, it was noted that studies have reported a relationship between grade levels and expected grade levels and SET scores. When this relationship is proactively pursued by teachers via a conscious easing up on grades and coursework, there appears to a kind of "mutual back patting" taking place where the teacher gives a high grade to the student (this grade not necessarily reflecting any real student attainment) and, in return, the student rewards the teacher with a high teacher rating. However, teacher influencing tactics need to be distinguished from what is termed in this paper and in the analytical framework, teacher behavioural traits which is another consistent area of research and can be summarized as the effect on SET scores of teacher "likeability". Finally, other teacher related aspects that have been explored by more than one study are the effect on SET scores of age, experience and rank. What follows is a more detailed description of the teacher related factors.

Gender. A great deal has been written about the affect of teachers' gender on SET results often on the premise that female teachers may be discriminated against in what may still be perceived of as a male dominated profession (Koh and Tan, 1997). However, studies of gender effects on SET results do not support a view that female teachers are consistently discriminated against. Thus, Bennett (1982) found that female teachers were consistently rated as friendlier, having a more positive interpersonal style and possessing greater charisma than their male counterparts. Similarly, female teachers have been rated higher than male teachers on the ability to create a classroom environment that invites participation (Crawford and MacLeod, 1990) and on the fostering of a feeling of closeness and warmth for both male and female students (Sears and Hennessey, 1996). Further, a meta-analysis of gender effect on student evaluations conducted by Feldman (1993) indicated that when significant differences were found, they generally favoured the female teacher.

Research also indicates that student ratings are strongly influenced by gender role expectations and, in general, it appears that teacher behaviour perceived by ratees to be inconsistent with traditional gender roles is penalized in student evaluations (Langbein, 1994). For example, females may be expected to be generally more caring and nurturing than men and if a female teacher does not display such qualities in the view of her students, she may well be penalized in her ratings. Similarly, males may be expected to be more directive and focused on the task than females and likewise may be penalized in student evaluations because students do not perceive them to be operating as expected. Rubin (1981), for instance, found that nurturing qualities were perceived of as more important for female professors than male professors and openness (fairness) more important for male professors. Similarly, Kierstead *et al.* (1988), in asking students to evaluate an imaginary teacher who was male in half the surveys and female in the other half, found that while warmth and interpersonal contact were viewed as important qualities for both male and female versions, the presence of these qualities only influenced students' evaluations of a notional female teacher. Equally, accessibility outside the classroom and a friendly attitude in the class (indicated by a regular smile) positively influenced evaluations of the imaginary female teacher and had no affect on ratings of the male version in the case of accessibility and, in the case of "the ready smile", reduced students' ratings of the male version.

In general, it appears that a number of traits such as warmth, charisma, accessibility, self-assurance and professionalism are valued across faculty gender (Bennett, 1982; Downs and Downs, 1993) but their influence on SET results tends to reflect gender stereotyping. Thus, female teachers perceived of as warm, charismatic and accessible are likely to be more positively evaluated on these traits than their male counterparts (Bennett, 1982; Cooper *et al.*, 1982, Kierstead *et al.*, 1988). Nevertheless, gender stereotyping of female teachers does not always produce positive results for them. Some studies have indicated that stereotyping may alert raters to a perceived shortcoming based on gender that might result in a severe rating if that shortcoming appears to be evident. Therefore, female teachers may be generally perceived to be less professional (professionalism being perceived of as a male quality) than their male colleagues and if the female teacher does not display such a high standard of professionalism that offsets the perception, the female teacher may incur a more negative rating than might otherwise have been the case (Bennett, 1982; Winocur *et al.*, 1989). In summary, the gender-student evaluation relationship is a complex but nonetheless significant factor influencing SETs.

Age, experience, rank. Smith and Kinney (1992) have suggested that the age of a teacher has an effect on SET scores and that older and more experienced teachers tend to receive more positive student evaluations. Furthermore, Holtfreter (1991) found a positive but weak relationship between the rank of a university teacher and student ratings. However, Feldman's (1983) comprehensive review of studies focusing of the influence of teachers' academic rank, instructional experience and age on SETs was not conclusive. Langbein (1994), on the other hand, did find a significant relationship between instructional experience and student ratings although this relationship was non-linear with experience having a positive effect on evaluations up to a point when the effect then became negative. Contrasting with the findings of Smith and Kinney (1992) and Holtfreter (1991), Clayson (1999) found that student evaluations tended to be negatively correlated with the teacher's age and years of experience. In summary, research has produced mixed results and indicates only a potential relationship between teacher age, experience and rank and student ratings.

Teachers' influencing tactics. Earlier, it was noted that despite the widespread use of the SET as the central measure of university teaching performance, academics have little confidence in its accuracy (Reckers, 1995). Furthermore, SET results often are a major input to personnel decisions relating to academic staff. This situation encourages university teachers to use various tactics to influence student evaluations, many of which, at best, have little educational value and at worst, are actually detrimental to the educational process. As one study suggests:

> This SET system causes professors to manipulate students and students in turn to manipulate teachers (Crumbley *et al.*, 2001).

Central to this manipulation are grades. A number of authors have noted that a common method used by teachers to court popularity is grade inflation and "easing up" on course content, assignments and tests (Bauer, 1996; Crumbley, 1995; Handlin, 1996; Ryan *et al.*, 1980; Sacks, 1996). To put it succinctly, university teachers can buy ratings with grades (Hocutt (1987-1988). In a review of faculty tactics aimed at influencing SET outcomes, Simpson and Siguaw (2000) found that the most significant factor reported by faculty was grading leniency and associated activities such as easy or no exams, unchallenging course material and spoon feeding students on examination content. In brief, many university teachers believe that lenient grading produces higher SET scores and they tend to act on this belief (Martin, 1998; Powell, 1977; Stumpf and Freedman, 1979; Winsor, 1977; Worthington and Wong, 1979; Yunker and Marlin, 1984).

Various other manipulative tactics are reported in the literature, many of them fatuous in an educational sense to say the least. For example, Emery (1995) found in a study of 2,673 students at a major US university that teachers who brought food to class received the highest ratings of teaching effectiveness. Simpson and Siguaw (2000) reported that university teachers perceived a major influencing tactic to be the serving of snacks etc. on the day of the evaluations. Other tactics noted by these authors included consistently letting students out of class early, complimenting the class on its ability immediately before administering the evaluation, administering the evaluation when poor students are absent, having a "fun activity" during the class on the day before the evaluation and remaining in the room during the evaluation. Not all the tactics noted by the authors were as irrelevant to the educational process. Some respondents stated that they provided their students with academic extras such as small, in-class, discussion groups and extra study sessions and others stated that they clearly outlined to their students what teaching and learning should be at university level and highlighted expectations in the syllabus. These academic extras were viewed as means of enhancing evaluations via improving students' academic performance and influencing student expectations. Despite these more positive approaches to influencing SET outcomes, it is evident that much of what is done by academics to influence student evaluations is of little or no educational value.

Teachers' behavioural traits. This section is distinguished from the previous section in concentrating on the influence of the more subtle university teachers' behaviour and character traits on SET's. This is very different from the above focus on the overt, sometimes cynical actions, used by some academics to positively influence SET results. Studies of the effect of personality variables on student evaluations are limited (Simpson and Siguaw, 2000). However, the research that has been done confirms that the behaviour traits of university teachers have a substantial impact on student evaluations. Thus, Feldman (1986) found that the overall relationship of teacher personality to student ratings is substantial. Williams and Ceci (1997) also found that student ratings are significantly influenced by the personal characteristics of the teacher. Similarly, Cardy and Dobbins (1986) found that students' "liking" of the teacher significantly influenced teaching evaluations. Clayson's (1999) study confirmed that between 50 per cent and 80 per cent of the total variance of student evaluations could be attributed to personality related variables. In a quantitative study, Jackson *et al.* (1999) found that a university teacher's ability to "get on" with students (rapport) overlapped heavily with more squarely educational factors such as teacher enthusiasm for subject, breadth of subject coverage, group interaction and learning value. An extreme interpretation of the type of findings reported by Jackson *et al.* (1999) would support Abrami *et al.* (1982) argument that personable faculty can receive favourable student ratings regardless of how well they know their subject matter (see, for example, Naflulin *et al.*, 1973). In sum, research indicates that university teachers' behavioural traits have a substantial effect on SET results. Studies have also suggested that these behavioural traits may not necessarily be of any educational value.

3. Conclusion and discussion

The above framework highlights the variety of factors influencing the accuracy of student evaluation of teaching and arguably encompasses the major research areas and themes. It is designed to help the researcher and practitioner make sense of the numerous studies that have focused on the SET phenomenon. Perusal of the factors contained in the framework indicates that, although the SET system has its advocates (see, for example, d'Apollonia and Abrami, 1997; Marsh and Roche, 1997; McKeachie, 1987), by and large, most studies have called into question the value of the SET system. It seems that there are so many variables unrelated to the actual execution of teaching influencing SET scores that they tend to obscure accurate assessment of teaching performance. Equally, SET research has generally failed to demonstrate that there is a concrete relationship between teaching performance and student achievement. Accordingly, analysis of the research using the framework presented in this paper suggests the time seems right to explore other methods of evaluating the quality of the classroom experience that could give a more accurate and comprehensive picture of classroom dynamics. For example, a recent study focused on classroom leadership, a notion broader than teaching, and found that effective classroom leadership stimulates extra effort among students (Pounder, 2005). The classroom leadership notion, for example, has considerable potential given the number of studies linking student effort and student achievement (Carbonaro, 2005; Eskew and Faley, 1988; Johnson *et al.*, 2002).

In conclusion, the title of this paper asked the following question: is student evaluation of teaching worthwhile? The framework presented here suggests that in the case of the SET process in its conventional form, its value is questionable as the sole measure of classroom performance since the quality, richness and diversity of what happens in the typical classroom cannot be captured by the SET process alone. However, in the field of education, measures of classroom effectiveness are essential despite the deficiencies of the conventional SET approach. There are therefore strong grounds for arguing that educational organizations can and should experiment with and develop approaches to assessing classroom dynamics that break from the conventional SET mold. Educational organizations might then be in the position to supplement the conventional SET with other approaches that have the potential to provide a richer picture, and more equitable assessment, of what happens in the classroom.

References

Abrami, P.C., Leventhal, L. and Perry, R.P. (1982), "Education seduction", *Review of Educational Research*, Vol. 32, pp. 446-64.

Aleamoni, L.M. (1981), "Student ratings of instruction", in Millman, J. (Ed.), *Handbook of Teacher Evaluation*, Sage Publications, Newbury Park, CA.

Aleamoni, L.M. (1989), "Typical faculty concerns about evaluation of teaching", in Aleamoni, L.M. (Ed.), *Techniques for Evaluating and Improving Instruction*, Jossey-Bass, San Francisco, CA.

Aronson, G. and Linder, D.E. (1965), "Gain and loss of esteem as determinants of interpersonal attractiveness", *Journal of Experimental Social Psychology*, Vol. 1, pp. 156-71.

Bachen, C.M., McLoughlin, M.M. and Garcia, S.S. (1999), "Assessing the role of gender in college students' evaluations of faculty", *Communication Education*, Vol. 48 No. 3, pp. 193-210.

Bauer, H.H. (1996), "The new generations: students who don't study", paper presented at The Technological Society at Risk Symposium, Orlando, FL.

Bennett, S.K. (1982), "Student perceptions of and expectations for male and female instructors: evidence relating to the question of gender bias in teaching evaluation", *Journal of Educational Psychology*, Vol. 74, pp. 170-9.

Brown, D.L. (1976), "Faculty ratings and student grades: a university-wide multiple regression analysis", *Journal of Educational Psychology*, Vol. 68 No. 5, pp. 573-8.

Carbonaro, W. (2005), "Tracking students' efforts, and academic achievement", *Sociology of Education*, Vol. 78 No. 1, pp. 27-49.

Cardy, R.L. and Dobbins, G.H. (1986), "Affect and appraisal accuracy: liking as an integral dimension in evaluating performance", *Journal of Applied Psychology*, Vol. 71, pp. 672-8.

Cashin, W. (1990), "Students do rate different academic fields differently", in Theall, M. and Franklin, J. (Eds), *Student Ratings of Instruction: Issues for Improving Practice*, Jossey-Bass, San Francisco, CA.

Centra, J.A. and Creech, F.R. (1976), *The Relationship Between Student, Teacher, and Course Characteristics and Student Ratings of Teacher Effectiveness. SIR Report No. 4*, Educational Testing Service, Princeton, NJ, pp. 24-7.

Clark, D. (1993), "Teacher evaluation: a review of the literature with implications for educators", Seminar in Elementary Education, California State University, Long Beach, CA, Spring.

Clayson, D.E. (1999), "Students' evaluation of teaching effectiveness: some implications of stability", *Journal of Marketing Education*, Vol. 21, April, pp. 68-75.

Cohen, P.A. (1983), "Comment on a selective review of the validity of student ratings of teaching", *Journal of Higher Education*, Vol. 54, pp. 448-58.

Cooper, P., Stewart, L. and Gudykunst, W.B. (1982), "Relationship with instructor and other variables influencing student evaluations of instruction", *Communication Quarterly*, Vol. 30, pp. 308-15.

Cranton, P. and Smith, R.A. (1986), "A new look at the effect of course characteristics on student ratings of instruction", *American Educational Research Journal*, Spring, pp. 117-28.

Crawford, M. and MacLeod, M. (1990), "Gender in the college classroom: an assessment of the 'chilly climate' for women", *Sex Roles*, Vol. 23, pp. 101-22.

Cronin, L. and Capie, W. (1986), "The influence of daily variation in teacher performance on the reliability and validity of assessment data", paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Crumbley, D.L. (1995), "The dysfunctional atmosphere of higher education: games professors play", *Accounting Perspectives*, Vol. 1 No. 1, pp. 27-33.

Crumbley, L., Henry, B.K. and Kratchman, S.H. (2001), "Students' perceptions of the evaluation of college teaching", *Quality Assurance in Education*, Vol. 9 No. 4, pp. 197-207.

Damron, J.C. (1996), *Instructor Personality and the Politics of the Classroom*, Douglas College, New Westminster.

d'Apollonia, S. and Abrami, P.C. (1997), "Navigating student ratings of instruction", *American Psychologist*, Vol. 52 No. 11, pp. 1198-208.

DeBerg, C.L. and Wilson, J.R. (1990), "An empirical investigation of the potential confounding variables in student evaluation of teaching", *Journal of Accounting Education*, Vol. 8 No. 1, pp. 37-62.

Dowell, D.A. and Neal, J.A. (1982), "A selective view of the validity of student ratings of teaching", *Journal of Higher Education*, No. 53, pp. 51-62.

Downs, V.C. and Downs, T.M. (1993), DC, "An exploratory and descriptive study identifying communicative behaviors associated with effective college teaching", paper presented at the Annual meeting of the International Communication Association, Washington, DC.

Emery, C.R. (1995), *Student Evaluations of Faculty Performance*, Clemson University, Clemson, SC.

Eskew, R.K. and Faley, R.H. (1988), "Some determinants of student performance in the first college-level financial accounting course", *The Accounting Review*, Vol. 63 No. 1, pp. 137-47.

Feldman, K.A. (1983), "Seniority and experience of college teachers as related to evaluations they receive from students", *Research in Higher Education*, Vol. 18 No. 1, pp. 3-124.

Feldman, K.A. (1984), "Class size and college students' evaluations of teachers and courses: a closer look", *Research in Higher Education*, Vol. 21 No. 1, pp. 45-116.

Feldman, K.A. (1986), "The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: a review and synthesis", *Research in Higher Education*, Vol. 24, pp. 139-213.

Feldman, K.A. (1993), "College students' views of male and female college teachers: Part II – evidence from students' evaluations of their classroom teachers", *Research in Higher Education*, Vol. 34, pp. 151-91.

Frey, P.W., Leonard, D.W. and Beatty, W.M. (1975), "Student ratings of instruction: validation research", *American Educational Research Journal*, Vol. 12 No. 4, pp. 435-47.

Glass, G.V., McGaw, B. and Smith, M.L. (1981), *Meta-Analysis in Social Research*, Sage, Beverly Hills, CA.

Goldman, L. (1993), "On the erosion of education and the eroding foundations of teacher education", *Teacher Education Quarterly*, Vol. 20, pp. 57-64.

Greenwald, A.G. (1997), "Validity concerns and usefulness of student ratings of instruction", *American Psychologist*, Vol. 52 No. 11, pp. 1182-7.

Handlin, O. (1996), "A career at Harvard", *American Scholar*, Vol. 65 No. 5, pp. 47-58.

Hocutt, M.O. (1987-1988), "De-grading student evaluations: what's wrong with student polls of teaching", *Academic Questions*, Winter, pp. 55-64.

Holtfreter, R.E. (1991), "Student rating biases: are faculty fears justified?", *The Woman CPA*, Fall, pp. 59-62.

Husbands, C.T. and Fosh, P. (1993), "Students' evaluation of teaching in higher education: experiences from four European countries and some implications of the practice", *Assessment and Evaluation in Higher Education*, Vol. 18 No. 2, pp. 95-114.

Jackson, D.L., Teal, C.R., Raines, S.J. and Nansel, T.R. (1999), "The dimensions of students' perceptions of teaching effectiveness", *Educational and Psychological Measurement*, Vol. 59 No. 4, pp. 580-96.

Johnson, D.L., Joyce, P. and Sen, S. (2002), "An analysis of student effort and performance in the finance principles course", *Journal of Applied Finance*, Vol. 12 No. 2, pp. 67-72.

Johnson, R.L. and Christian, V.K. (1990), "Relation of perceived learning and expected grade to rated effectiveness of teaching", *Perceptual and Motor Skills*, Vol. 70, pp. 479-82.

Kierstead, D., D'Agostino, P. and Dill, H. (1988), "Sex role stereotyping of college professors: bias in student ratings of instructors", *Journal of Educational Psychology*, Vol. 80 No. 3, pp. 342-4.

Koh, C.H. and Tan, T.M. (1997), "Empirical investigation of the factors affecting SET results", *International Journal of Educational Management*, Vol. 11 No. 4, pp. 170-8.

Langbein, L.I. (1994), "The validity of student evaluations of teaching", *Political Science and Politics*, September, pp. 545-53.

Liaw, S.H. and Goh, K.L. (2003), "Evidence and control of biases in student evaluations of teaching", *The International Journal of Educational Management*, Vol. 17 No. 1, pp. 37-43.

McCallum, L.W. (1984), "A meta-analysis of course evaluation data and its use in the tenure decision", *Research in Higher Education*, Vol. 21, pp. 150-8.

McKeachie, W. (1987), "Can evaluating instruction improve teaching?", in Aleamoni, L.M. (Ed.), *Techniques for Evaluating and Improving Instruction*, Jossey-Bass, San Francisco, CA.

Marsh, H.W. (1984), "Students' evaluation of university teaching: dimensionality, reliability, validity, potential biases, and utility", *Journal of Educational Psychology*, October, pp. 707-54.

Marsh, H.W. (1987), "Students' evaluations of university teaching: research findings, methodological issues, and directions for future research", *Journal of Educational Research*, Vol. 11, pp. 253-388.

Marsh, H.W. and Roche, L.A. (1997), "Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias and utility", *American Psychologist*, Vol. 52 No. 11, pp. 1187-97.

Martin, J.R. (1998), "Evaluating faculty based on student opinions: problems, implications and recommendations from Deming's theory of management perspective", *Issues in Accounting Education*, Vol. 13 No. 4, pp. 1079-94.

Meredith, G.M. (1984), "Diagnostic and summative appraisal ratings of instruction", *Psychological Reports*, Vol. 46, pp. 21-2.

Naflulin, D., Ware, J. and Donnelly, F. (1973), "The Dr Fox lecture: a paradigm of educational seduction", *Journal of Medical Education*, Vol. 48, pp. 630-5.

Newton, J.D. (1988), "Using student evaluation of teaching in administrative control: the validity problem", *Journal of Accounting Education*, Vol. 6 No. 1, pp. 1-14.

Perkins, D., Gueri, D. and Schleh, J. (1990), "Effects of grading standards information, assigned grade, and grade discrepancies on student evaluations", *Psychological Reports*, Vol. 66, pp. 635-42.

Pounder, J.S. (2005), "The classroom leadership styles of Hong Kong university teachers: a case study of teachers in a business school", doctoral dissertation, School of Education, Centre for Educational Leadership and Management, University of Leicester.

Powell, R.W. (1977), "Grades, learning, and student evaluation of instruction", *Research in Higher Education*, Vol. 7, pp. 193-205.

Reckers, P.M.J. (1995), "Know thy customer", in Baril, C.P. (Ed.), *Change in Accounting Education: A Research Blueprint*, Federation of Schools of Accountancy, St Louis, MO.

Rubin, R.B. (1981), "Ideal traits and terms of address for male and female college professors", *Journal of Personality and Social Psychology*, Vol. 41, pp. 966-74.

Ryan, J.I., Anderson, J.A. and Birchler, A.B. (1980), "Evaluations: the faculty responds", *Research in Higher Education*, Vol. 12 No. 4, pp. 317-33.

Sacks, P. (1996), *Generation X Goes to College*, Open Court, Chicago, IL.

Sears, S.R. and Hennessey, A.C. (1996), "Students' perceived closeness to professors: the effects of school, professor gender and student gender", *Sex Roles*, Vol. 35, pp. 651-8.

Seldin, P. (1984), *Changing Practices in Faculty Evaluation*, Jossey-Bass, San Francisco, CA.

Seldin, P. (1989), "How colleges evaluate professors", *American Association for Higher Education Bulletin*, Vol. 41 No. 7, pp. 3-7.

Seldin, P. (1993), "The use and abuse of student ratings of professors", *The Chronicle of Higher Education*, Vol. 39 No. 46, p. A40.

Simpson, P.M. and Siguaw, J.A. (2000), "Student evaluations of teaching: an exploratory study of the faculty response", *Journal of Marketing Education*, Vol. 22 No. 3, pp. 199-213.

Smith, S.P. and Kinney, D.P. (1992), "Age and teaching performance", *Journal of Higher Education*, Vol. 63 No. 3, pp. 282-302.

Stodolsky, S. (1984), "Teacher evaluation: the limits of looking", *Educational Researcher*, November, pp. 11-18.

Stratton, W.O. (1990), "A model for the assessment of student evaluations of teaching, and the professional development of faculty", *The Accounting Educators' Journal*, Summer, pp. 77-101.

Stumpf, S.A. and Freedman, R.D. (1979), "Expected grade covariation with student ratings of instructors", *Journal of Educational Psychology*, Vol. 71, pp. 273-302.

Toby, S. (1993), "Class size and teaching evaluation", *Journal of Chemical Education*, Vol. 70 No. 6, pp. 465-6.

Walumbwa, F.O. and Ojode, L.A. (2000), "Gender stereotype and instructors' leadership behavior: transformational and transactional leadership", paper presented at the Midwest Academy of Management Annual Conference, Chicago, IL, 30 March-1 April.

Williams, W.M. and Ceci, S.J. (1997), "'How'm I doing?' Problems with student ratings of instructors and courses", *Change*, Vol. 29 No. 5, pp. 12-23.

Wilson, R. (1998), "New research casts doubt on value of student evaluations of professors", *The Chronicle of Higher Education*, Vol. 44 No. 19, pp. A12-A14.

Winocur, S., Schoen, L.G. and Sirowatka, A.H. (1989), "Perceptions of male and female academics within a teaching context", *Research in Higher Education*, Vol. 30, pp. 317-29.

Winsor, J.L. (1977), "A's, B's, but not C's: a comment", *Contemporary Education*, Vol. 48, pp. 82-4.

Worthington, A.G. and Wong, P.T.P. (1979), "Effects of earned and assigned grades on student evaluations of an instructor", *Journal of Educational Psychology*, Vol. 71, pp. 764-75.

Yunker, J.A. and Marlin, J.W. (1984), "Performance evaluation of college and university faculty: an economic perspective", *Educational Administration Quarterly*, Winter, pp. 9-37.

Further reading

Gellis, Z.D. (2001), "Social work perceptions of transformational and transactional leadership in health care", *Social Work Research*, Vol. 25 No. 1, pp. 17-25.

Michaels, J.W. and Miethe, T.D. (1989), "Academic effort and college grades", *Social Forces*, Vol. 68 No. 1, pp. 309-19.

Naser, K. and Peel, M.J. (1996), "An exploratory study of the impact of intervening variables on student performance in a principles of accounting course", *Accounting Education*, Vol. 7 No. 3, pp. 209-23.

Williams, R.L. and Clark, L. (2002), *Academic Causal Attribution and Course Outcomes for College Students*, ERIC ED 469 337.

This article has been cited by:

1. Twan Huybers, Jordan Louviere, Towhidul Islam. 2015. What determines student satisfaction with university subjects? A choice-based approach. *Journal of Choice Modelling* . [CrossRef]
2. James S. Pounder, Elizabeth Ho Hung-lam, Julie May Groves. 2015. Faculty-student engagement in teaching observation and assessment: a Hong Kong initiative. *Assessment & Evaluation in Higher Education* 1-13. [CrossRef]
3. Chenicheri Sid Nair, Jinrui Li, Li Kun Cai. 2015. Academics' feedback on the quality of appraisal evidence. *Quality Assurance in Education* **23**:3, 279-294. [Abstract] [Full Text] [PDF]
4. 2014. An Evaluation of Course Evaluations. *ScienceOpen Research* . [CrossRef]
5. Adriana Morales Rodríguez, Joan-Lluís Capelleras, Víctor M. Gimenez Garcia. 2014. Teaching performance: determinants of the student assessment. *Academia Revista Latinoamericana de Administración* **27**:3, 402-418. [Abstract] [Full Text] [PDF]
6. Salochana Hassan, Wouter Wium. 2014. Quality lies in the eyes of the beholder: A mismatch between student evaluation and peer observation of teaching. *Africa Education Review* **11**, 491-511. [CrossRef]
7. Kudzayi Maumbe. 2014. Teaching and Learning in Recreation and Tourism: A Comparison of Three Instructional Methods. *Journal of Teaching in Travel & Tourism* **14**, 365-385. [CrossRef]
8. Twan Huybers. 2014. Student evaluation of teaching: the use of best–worst scaling. *Assessment & Evaluation in Higher Education* **39**, 496-513. [CrossRef]
9. Joseph F. Fletcher, Michael A. Painter-Main. 2014. An Elephant in the Room: Bias in Evaluating a Required Quantitative Methods Course. *Journal of Political Science Education* **10**, 121-135. [CrossRef]
10. Donald Morley. 2014. Assessing the reliability of student evaluations of teaching: choosing the right coefficient. *Assessment & Evaluation in Higher Education* **39**, 127-139. [CrossRef]
11. Satish Nargundkar, Milind Shrikhande. 2014. Norming of Student Evaluations of Instruction: Impact of Noninstructional Factors. *Decision Sciences Journal of Innovative Education* **12**:10.1111/dsji.2014.12.issue-1, 55-72. [CrossRef]
12. Hamzeh Dodeen. 2013. Validity, Reliability, and Potential Bias of Short Forms of Students' Evaluation of Teaching: The Case of UAE University. *Educational Assessment* **18**, 235-250. [CrossRef]
13. Keith Morrison, Trav Johnson. 2013. EDITORIAL. *Educational Research and Evaluation* **19**, 579-584. [CrossRef]
14. Beatrice Tucker, Beverley Oliver, Ritu Gupta. 2013. Validating a teaching survey which drives increased response rates in a unit survey. *Teaching in Higher Education* **18**, 427-439. [CrossRef]
15. Lyn Alderman, Stephen Towers, Sylvia Bannah. 2012. Student feedback systems in higher education: a focused literature review and environmental scan. *Quality in Higher Education* **18**, 261-280. [CrossRef]
16. Stuart Palmer. 2012. Student evaluation of teaching: keeping in touch with reality. *Quality in Higher Education* **18**, 297-311. [CrossRef]
17. Stephen Darwin. 2012. Moving beyond face value: re-envisioning higher education evaluation as a generator of professional knowledge. *Assessment & Evaluation in Higher Education* **37**, 733-745. [CrossRef]
18. Stephen L. Wright, Michael A. Jenkins-Guarnieri. 2012. Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education* **37**, 683-699. [CrossRef]

19. Craig S. Galbraith, Gregory B. Merrill. 2012. Faculty Research Productivity and Standardized Student Learning Outcomes in a University Teaching Environment: A Bayesian Analysis of Relationships. *Studies in Higher Education* **37**, 469-480. [CrossRef]

20. Craig S. Galbraith, Gregory B. Merrill, Doug M. Kline. 2012. Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses. *Research in Higher Education* **53**, 353-374. [CrossRef]

21. Marta Barandiaran-Galdós, Miren Barrenetxea Ayesta, Antonio Cardona-Rodríguez, Juan José Mijangos del Campo, Jon Olaskoaga-Larrauri. 2012. What do teachers think about quality in the Spanish university?. *Quality Assurance in Education* **20**:2, 91-109. [Abstract] [Full Text] [PDF]

22. Craig S. Galbraith, Gregory B. Merrill. 2012. Predicting Student Achievement in University-Level Business and Economics Classes: Peer Observation of Classroom Instruction and Student Ratings of Teaching Effectiveness. *College Teaching* **60**, 48-55. [CrossRef]

23. Seyedeh Azadeh Safavi, Kamariah Abu Bakar, Rohani Ahmad Tarmizi, Nor Hayati Alwi. 2012. What do higher education instructors consider useful regarding student ratings of instruction? Limitations and recommendations. *Procedia - Social and Behavioral Sciences* **31**, 653-657. [CrossRef]

24. Dennis C.S. Law, Jan H.F. Meyer. 2011. Initial investigation of Hong Kong post-secondary students' learning patterns. *Quality Assurance in Education* **19**:4, 335-356. [Abstract] [Full Text] [PDF]

25. Stephen Wilkins, Alun Epps. 2011. Student evaluation web sites as potential sources of consumer information in the United Arab Emirates. *International Journal of Educational Management* **25**:5, 410-422. [Abstract] [Full Text] [PDF]

26. John Rogers, Morgan Smith. 2011. Demonstrating genuine interest in students' needs and progress. *Journal of Applied Research in Higher Education* **3**:1, 6-14. [Abstract] [Full Text] [PDF]

27. Stephen Wilkins. 2010. Higher education in the United Arab Emirates: an analysis of the outcomes of significant increases in supply and competition. *Journal of Higher Education Policy and Management* **32**, 389-400. [CrossRef]

28. GUO-HAI CHEN, DAVID WATKINS. 2010. CAN STUDENT RATINGS OF TEACHING BE PREDICTED BY TEACHING STYLES? 1. *Psychological Reports* **106**, 501-512. [CrossRef]

29. Thorsten Gruber, Stefan Fuß, Roediger Voss, Michaela Gläser-Zikuda. 2010. Examining student satisfaction with higher education services. *International Journal of Public Sector Management* **23**:2, 105-123. [Abstract] [Full Text] [PDF]

Economics of Education Review 41 (2014) 71–88

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Evaluating students' evaluations of professors[☆]

Michela Braga [a], Marco Paccagnella [b], Michele Pellizzari [c,*]

[a] Bocconi University, Department of Economics, Italy
[b] Bank of Italy, Trento Branch, Italy
[c] University of Geneva, Institute of Economics and Econometrics, Switzerland

ARTICLE INFO

Article history:
Received 2 August 2013
Received in revised form 22 April 2014
Accepted 23 April 2014
Available online 5 May 2014

JEL classification:
I20
M55

Keywords:
Teacher quality
Postsecondary education
Students' evaluations

ABSTRACT

This paper contrasts measures of teacher effectiveness with the students' evaluations for the same teachers using administrative data from Bocconi University. The effectiveness measures are estimated by comparing the performance in follow-on coursework of students who are randomly assigned to teachers. We find that teacher quality matters substantially and that our measure of effectiveness is negatively correlated with the students' evaluations of professors. A simple theory rationalizes this result under the assumption that students evaluate professors based on their realized utility, an assumption that is supported by additional evidence that the evaluations respond to meteorological conditions.

1. Introduction

The use of anonymous students' evaluations of professors to measure teachers' performance has become extremely popular in many universities (Becker & Watts, 1999). They normally include questions about the clarity of lectures, the logistics of the course, and many others. They are either administered during a teaching session toward the end of the term or, more recently, filled on-line.

The university administration uses such evaluations to solve the agency problems related to the selection and motivation of teachers, in a context in which neither the types of teachers, nor their effort, can be observed precisely. In fact, students' evaluations are often used to inform hiring and promotion decisions (Becker & Watts, 1999) and, in institutions that put a strong emphasis on research, to avoid strategic behavior in the allocation of time or effort between teaching and research activities (Brown & Saks, 1987; De Philippis, 2013).[1]

[☆] We would like to thank Bocconi University for granting access to its administrative archives for this project. In particular, the following persons provided invaluable and generous help: Giacomo Carrai, Mariele Chirulli, Mariapia Chisari, Alessandro Ciarlo, Alessandra Gadioli, Roberto Grassi, Enrica Greggio, Gabriella Maggioni, Erika Palazzo, Giovanni Pavese, Cherubino Profeta, Alessandra Startari and Mariangela Vago. We are also indebted to Tito Boeri, Giovanni Bruno, Giacomo De Giorgi, Marco Leonardi, Vincenzo Mariani, Tommaso Monacelli, Tommy Murphy and Tommaso Nannicini for their precious comments. We would also like to thank seminar participants at the Bank of Italy, Bocconi University, International Workshop on Applied Economics of Education, London School of Economics, UC Berkeley, Università Statale di Milano and LUISS University. Davide Malacrino and Alessandro Ferrari provided excellent research assistance. Michele Pellizzari is also affiliated to IZA, CREAM, fRDB and NCCR-LIVES. The views expressed in this paper are solely those of the authors and do not involve the responsibility of the Bank of Italy. The usual disclaimer applies.

* Corresponding author at: University of Geneva, Department of Economics, Uni-Mail, 40 Bd. du Pont d'Arve, CH-1211 Geneva 4, Switzerland. Tel.: +41 22 37 98 915.
E-mail addresses: michele.pellizzari@unige.ch (M. Pellizzari).

[1] Although there is some evidence that a more research oriented faculty also improve academic and labor market outcomes of graduate students (Hogan, 1981).

http://dx.doi.org/10.1016/j.econedurev.2014.04.002
0272-7757/© 2014 Elsevier Ltd. All rights reserved.

The validity of anonymous students' evaluations rests on the assumption that, by attending lectures, students observe the ability of the teachers and that they report it truthfully when asked. While this view is certainly plausible, there are also many reasons to question the appropriateness of such a measure. For example, the students' objectives might be different from those of the principal, i.e. the university administration. Students may simply care about their grades, whereas the university cares about their learning and the two might not be perfectly correlated, especially when the same professor is engaged both in teaching and in grading. Consistent with this interpretation, Krautmann and Sander (1999) show that, conditional on learning, teachers who give higher grades also receive better evaluations. This finding is confirmed by several other studies and is thought to be a key cause of grade inflation (Carrell & West, 2010; Johnson, 2003; Weinberg, Fleisher, & Hashimoto, 2009).

Measuring teaching quality is complicated also because the most common observable teachers' characteristics, such as qualifications or experience, appear to be relatively unimportant (Hanushek & Rivkin, 2006; Krueger, 1999; Rivkin, Hanushek, & Kain, 2005). Despite such difficulties, there is evidence that teachers' quality matters substantially in determining students' achievement (Carrell & West, 2010; Rivkin et al., 2005) and that teachers respond to incentives (Duflo, Hanna, & Ryan, 2012; Figlio & Kenny, 2007; Lavy, 2009). Hence, understanding how professors should be monitored and incentivized is essential for education policy.

In this paper we evaluate the content of the students' evaluations by contrasting them with objective measures of teacher effectiveness. We construct such measures by comparing the performance in subsequent coursework of students who are randomly allocated to different teachers in their compulsory courses. We use data about one cohort of students at Bocconi University – the 1998/1999 freshmen – who were required to take a fixed sequence of compulsory courses and who where randomly allocated to a set of teachers for each of such courses.

We find that, even in a setting where the syllabuses are fixed and all teachers in the same course present exactly the same material, professors still matter substantially. The average difference in subsequent performance between students assigned to the best and worst teacher (on the effectiveness scale) is approximately 23% of a standard deviation in the distribution of exam grades, corresponding to about 3% of the average grade. Moreover, our measure of teaching quality is negatively correlated with the students' evaluations of the professors: teachers who are associated with better subsequent performance receive worst evaluations from their students. On the other hand, teachers who are associated with high grades in their own exams rank higher in the students' evaluations.

These results question the idea that students observe the ability of the teacher during the class and report it (truthfully) in their evaluations. In order to rationalize our findings it is useful to think of good teachers – i.e. those who provide their students with knowledge that is useful in future learning – as teachers who require effort from their students. Students dislike exerting effort, especially the least able ones, and when asked to evaluate the teacher they do so on the basis of how much they enjoyed the course. As a consequence, good teachers can get bad evaluations, especially if they teach classes with a lot of bad students.

Consistent with this intuition, we also find that the evaluations of classes in which high-skill students are over-represented are more in line with the estimated quality of the teacher. Additionally, in order to provide evidence supporting the intuition that evaluations are based on students' realized utility, we collected data on the weather conditions observed on the exact days when students filled the questionnaires. Assuming that the weather affects utility and not teaching quality, the finding that the students' evaluations react to meteorological conditions lends support to our intuition.[2] Our results show that students evaluate professors more negatively on rainy and cold days.

There is a large literature that investigates the role of teacher quality and teacher incentives in improving educational outcomes, although most of the existing studies focus on primary and secondary schooling (Figlio & Kenny, 2007; Jacob & Lefgren, 2008; Kane & Staiger, 2008; Rivkin et al., 2005; Rockoff, 2004; Rockoff & Speroni, 2010; Tyler, Taylor, Kane, & Wooten, 2010). The availability of internationally standardized test scores facilitates the evaluation of teachers in primary and secondary schools (Mullis, Martin, Robitaille, & Foy, 2009; OECD, 2010). The large degree of heterogeneity in subjects and syllabuses in universities makes it very difficult to design common tests that would allow to compare the performance of students exposed to different teachers, especially across subjects. At the same time, the large increase in college enrollment occurred in the past decades (OECD, 2008) calls for a specific focus on higher education.

Only very few papers investigate the role of students' evaluations in university and we improve on existing studies in various dimensions. First of all, the random allocation of students to teachers differentiates our approach from most other studies (Beleche, Fairris, & Marks, 2012; Johnson, 2003; Krautmann & Sander, 1999; Weinberg et al., 2009; Yunker & Yunker, 2003) that cannot purge their estimates from the potential bias due to the best students selecting the courses of the best professors. Correcting this bias is pivotal to producing reliable measures of teaching quality (Rothstein, 2009, 2010).

The only other study that exploits a setting where students are randomly allocated to teachers is Carrell and West (2010). This paper documents (as we do) a negative correlation between the students' evaluations of professors and harder measures of teaching quality. We improve on their analysis in two important dimensions. First, we provide additional empirical evidence consistent with an interpretation of such finding based on the idea that good professors require students to exert more effort and that students evaluate professors on the basis of their realized utility. Secondly, Carrell and West (2010) use data from a U.S. Air Force Academy, while our empirical application is

[2] One may actually think that also the mood of the professors, hence, their effectiveness in teaching is affected by the weather. However, students are asked to evaluate teachers' performance over the entire duration of the course and not exclusively on the day of the test. Moreover, it is a clear rule of the university to have students fill the questionnaires before the lecture, so that the teachers' performance on that specific day should not affect the evaluations.

based on a more standard institution of higher education.[3] The vast majority of the students in our sample enter a standard labor market upon graduation, whereas the cadets in Carrell and West (2010) are required to serve as officers in the U.S. Air Force for 5 years after graduation and many pursue a longer military career. There are many reasons why the behaviors of both teachers, students and the university/academy might vary depending on the labor market they face. For example, students may put higher effort on subjects or activities particularly important in the military setting at the expenses of other subjects and teachers and administrators may do the same.

More generally, this paper is also related and contributes to the wider literature on performance measurement and performance pay. One concern with the students' evaluations of teachers is that they might divert professors from activities that have a higher learning content for the students (but that are more demanding in terms of students' effort) and concentrate more on classroom entertainment (popularity contests) or change their grading policies. This interpretation is consistent with the view that teaching is a multi-tasking job, which makes the agency problem more difficult to solve (Holmstrom & Milgrom, 1994). Subjective evaluations can be seen as a mean to address such a problem and, given the very limited extant empirical evidence (Baker, Gibbons, & Murphy, 1994; Prendergast & Topel, 1996), our results can certainly inform also this area of the literature.

The paper is organized as follows. Section 2 describes the data and the institutional setting. Section 3 presents our strategy to estimate teacher effectiveness and shows the results. In Section 4 we correlate teacher effectiveness with the students' evaluations of professors. Robustness checks are reported in Section 5. In Section 6 we discuss the interpretation of our results and we present additional evidence supporting such an interpretation. Finally, Section 7 concludes.

2. Data and institutional details

The empirical analysis is based on data for one enrollment cohort of undergraduate students at Bocconi University, an Italian private institution of tertiary education offering degree programs in economics, management, public policy and law. We select the cohort of the 1998/1999 freshmen because it is the only one available where students were randomly allocated to teaching classes for each of their compulsory courses.[4]

The students entering Bocconi in the 1998/1999 academic year were offered 7 different degree programs but only three of them attracted enough students to require the splitting of lectures into more than one class: Management, Economics and Law&Management.[5] Students in these programs were required to take a fixed sequence of compulsory courses that span over the first two years, a good part of their third year and, in a few cases, also their last year. Table A.1 lists the exact sequences of the three programs.[6] We construct measures of teacher effectiveness for all and only the professors of these compulsory courses. We do not consider elective subjects, as the endogenous self-selection of students would complicate the analysis.

Most of the courses listed in Table A.1 were taught in multiple classes (see Section 3 for details). The number of classes varied across both degree programs and courses depending on the number of students and faculty. For example, Management was the program that attracted the most students (over 70% in our cohort), who were normally divided into 8–10 classes. Regardless of the class to which students were allocated, they were all taught the same material, with some variations across degree programs.

The exam questions were also the same for all students (within degree program), regardless of their classes. Specifically, one of the teachers in each course (normally a senior faculty member) acted as a coordinator, making sure that all classes progressed similarly during the term and addressing problems that might have arisen. The coordinator also prepared the exam paper, which was administered to all classes. Grading was delegated to the individual teachers, each of them marking the papers of the students in his/her own class. The coordinator would check that the distributions were similar across classes but grades were not curved, neither across nor within classes.

Our data cover the entire academic history of the students, including basic demographics, high school type and grades, the test score obtained in the cognitive admission test to the university, tuition fees (that varied with family income) and the grades in all exams they sat at Bocconi; graduation marks are observed for all non-dropout students.[7] Importantly, we also have access to the

[3] Bocconi is a selective college that offers majors in the wide area of economics, management, public policy and law, hence it is likely comparable to US colleges in the mid-upper part of the quality distribution. Faculty in the economics department hold PhDs from Harvard, MIT, NYU, Stanford, UCLA, LSE, Pompeu Fabra, Stockholm University. Also, the Bocconi Business school is ranked in the same range as the Georgetown University McDonough School of Business or the Johnson School at Cornell University in the US and to the Manchester Business School or the Warwick Business School in the UK (see the *Financial Times Business Schools Rankings*).

[4] The terms *class* and *lecture* often have different meanings in different countries and sometimes also in different schools within the same country. In most British universities, for example, *lecture* indicates a teaching session where an instructor – typically a full faculty member – presents the main material of the course; *classes* are instead practical sessions where a teacher assistant solves problem sets and applied exercises with the students. At Bocconi there was no such distinction, meaning that the same randomly allocated groups were kept for both regular lectures and applied classes. Hence, in the remainder of the paper we use the two terms interchangeably.

[5] The other degree programs were Economics and Social Disciplines, Economics and Finance, Economics and Public Administration.

[6] Notice that Economics and Management share exactly the same sequence of compulsory courses in the first three terms. Indeed, students in these two programs did attend these courses together and made a final decision about their major at the end of the third term. De Giorgi, Pellizzari, and Redaelli (2010) study precisely this choice. In the rest of the paper we abstract from this issue and we treat the two degree programs as entirely separated. Considering the two programs jointly does not affect our findings.

[7] The dropout rate, defined as the number of students who, according to our data, do not appear to have completed their programs at Bocconi over the total size of the entering cohort, is just above 10%. Notice that some of these students might have transferred to another university or still be working towards the completion of their program, whose formal duration was 4 years. Excluding the dropouts from our calculations is irrelevant for our results.

Table 1
Descriptive statistics of students.

Variable	Management	Economics	Law&Management	Total
1 = female	0.408	0.427	0.523	0.427
1 = outside Milan[a]	0.620	0.748	0.621	0.634
1 = top Income Bracket[b]	0.239	0.153	0.368	0.248
1 = academic high school[c]	0.779	0.794	0.684	0.767
1 = late enrollee[d]	0.014	0.015	0.011	0.014
High-school grade (0–100)	86.152	93.053	88.084	87.181
	(10.905)	(8.878)	(10.852)	(10.904)
Entry test score (0–100)	60.422	63.127	58.894	60.496
	(13.069)	(15.096)	(12.262)	(13.224)
University grades (0–30)	25.684	27.032	25.618	25.799
	(3.382)	(2.938)	(3.473)	(3.379)
Class size	121.29	28.55	125.28	127.12
	(62.20)	(33.00)	(44.14)	(62.84)
Number of students	901	131	174	1206

[a] Dummy equal to one if the student's place of residence at the time of first enrollment is outside the province of Milan (which is where Bocconi University is located).
[b] Family income is recorded in brackets and the dummy is equal to one for students who report incomes in the top bracket, whose lower threshold is in the order of approximately 110,000 euros at current prices.
[c] Dummy equal to one if the student attended a academic high school, such as a lyceum, rather than professional or vocational schools.
[d] Dummy equal to one if the student enrolled at Bocconi after age 19.

random class identifiers that allow us to identify in which class each students attended each of their courses.

Table 1 reports some descriptive statistics for the students in our data by degree program. The vast majority of them were enrolled in the Management program (74%), while Economics and Law&Management attracted 11% and 14%. Female students were under-represented in the student body (43% overall), apart from the degree program in Law&Management. Family income was recorded in brackets and one quarter of the students were in the top bracket, whose lower threshold was in the order of approximately 110,000 euros at current prices. Students from such a wealthy background were under-represented in the Economics program and over-represented in Law&-Management. High school grades and entry test scores (both normalized on the scale 0–100) provide a measure of ability and suggest that Economics attracted the best students.

Finally, we complement our dataset with students' evaluations of teachers. Towards the end of each term (typically in the last week), students in all classes were asked to fill an evaluation questionnaire during one lecture. The questions gathered students' opinions about various aspects of the teaching experience, including the clarity of the lectures, the logistics of the course, the availability of the professor. For each item in the questionnaire, students answered on a scale from 0 (very negative) to 10 (very positive) or 1 to 5.

In order to allow students to evaluate their experience without fear of retaliation from the teachers at the exam, the questionnaires were anonymous and it is impossible to match the individual student with a specific evaluation. However, each questionnaire reports the name of the course and the class identifier, so that we can attach average evaluations to each class in each course.

In Table 2 we present some descriptive statistics of the evaluation questionnaires. We concentrate on a limited set of items, namely overall teaching quality, lecturing clarity,

Table 2
Descriptive statistics of students' evaluations.

Variable	Management	Economics	Law&Manag.	Total
	Mean	Mean	Mean	Mean
	(std. dev.)	(std. dev.)	(std. dev.)	(std. dev.)
Overall teaching quality[a]	7.103	7.161	6.999	7.115
	(0.956)	(0.754)	(1.048)	(0.900)
Lecturing clarity[b]	3.772	3.810	3.683	3.779
	(0.476)	(0.423)	(0.599)	(0.467)
Teacher generates interest[a]	6.800	6.981	6.915	6.864
	(0.905)	(0.689)	(1.208)	(0.865)
Course logistic[b]	3.683	3.641	3.617	3.666
	(0.306)	(0.266)	(0.441)	(0.303)
Course workload[b]	2.709	2.630	2.887	2.695
	(0.461)	(0.542)	(0.518)	(0.493)
Response rate[c]	0.777	0.774	0.864	0.782
	(0.377)	(0.411)	(0.310)	(0.383)

See Table A.2 for the exact wording of the evaluation questions.
[a] Scores range from 0 to 10.
[b] Scores range from 1 to 5.
[c] Number of collected valid questionnaires over the number of officially enrolled students.

the teacher's ability to generate interest in the subject, the logistic of the course and workload.[8]

The average evaluation of overall teaching quality is around 7, with a relatively large standard deviation of 0.9 and minor variations across degree programs. Although differences are not statistically significant, professors in the Economics program seem to receive slightly better students' evaluations.

One might actually be worried that students may drop out of a class in response to the quality of the teaching so that at the end of the course, when questionnaires are distributed, only the students who liked the teacher are eventually present. Such a process would lead to a compression of the distribution of the evaluations, with good teachers being evaluated by their entire class (or by a majority of their allocated students) and bad teachers being evaluated only by a subset of students who particularly liked them.

The descriptive statistics reported in Table 2 seem to indicate that this is not a major concern, as on average the number of collected questionnaires is around 80% of the total number of enrolled students (the median is very similar). Moreover, when we correlate our measures of teaching effectiveness with the evaluations we condition on the official size of the class and we weight observations by the number of collected questionnaires. Indirectly, the relatively high response rate provides evidence that attendance was also pretty high. An alternative measure of attendance can be extracted from a direct question of the evaluation forms which asks students what percentage of the lectures they attended. Such self-reported measure of attendance is also around 80%.

Although the institutional setting at Bocconi-which facilitates this study-is somewhat unique, there is nothing about it that suggests these findings will not generalize; the consistency of our findings with other previous studies (Carrell & West, 2010; Krautmann & Sander, 1999; Weinberg et al., 2009) further implies that our analysis is not picking up something unique to Bocconi.

2.1. The random allocation

In this section we present evidence that the random allocation of students into classes was successful.[9] The randomization was (and still is) performed via a simple random algorithm that assign a class identifier to each student, who were then instructed to attend the lectures for the specific course in the class labeled with the same identifier. The university administration adopted the policy of repeating the randomization for each course with the explicit purpose of encouraging wide interactions among the students.

Table 3 is based on test statistics derived from probit (columns 1, 2, and 5–7) or OLS (columns 3 and 4) regressions of the observable students' characteristics (by column) on class dummies for each course in each degree program considered. The null hypothesis under consideration is the joint significance of the coefficients on the class dummies in each model, which amounts to testing for the equality of the means of the observable variables across classes. Considering descriptive statistics about the distribution of *p*-values for such tests, we observe that mean and median *p*-values are in all cases far from the conventional thresholds of 5% or 1% and only in a very few instances the null can be rejected. Overall the randomization was rather successful. Also the distributions of the available measures of students ability (high school grades and entry test scores) for the entire student body and for a randomly selected class in each program are extremely similar (see Fig. A.1).

Even though students were randomly assigned to classes, one may still be concerned about teachers being selectively allocated to classes. Although no explicit random algorithm was used to assign professors to classes, for organizational reasons the assignment process was done in the Spring of the previous academic year, well before students were allowed to enroll; therefore, even if teachers could choose their class identifiers they would have no chance to know in advance the characteristics of the students who would be given that same identifiers.

More specifically, there used to be a very strong hysteresis in the matching of professors to class identifiers, so that, if no particular changes occurred, one kept the same class identifier of the previous academic year: modifications took place only when some teachers needed to be replaced or the overall number of classes changed. Even in these instances, though, the distribution of class identifiers across professors changed only marginally. For example, if one teacher dropped out, then a new teacher would take his/her class identifier and none of the others were given a different one. Hence, most teachers maintain the same identifier over the years, provided they keep teaching the same course.[10]

About around the same time when teachers were given class identifiers, also classrooms and time schedules were defined. On these two items, though, teachers did have some limited choice. Typically, the administration suggested a time schedule and room allocation and professors could request one or more modifications, which were accommodated only if compatible with the overall teaching schedule.

In order to avoid any distortion in our estimates of teaching effectiveness due to the more or less convenient teaching times, we collected detailed information about the exact timing of the lectures in all the classes that we consider, so that we can hold this specific factor constant. Additionally, we also know in which exact room each class was taught and we further condition on the characteristics of the classrooms, namely the building and the floor where

[8] The exact wording and scaling of the questions are reported in Table A.2.

[9] De Giorgi et al. (2010) use data for the same cohort (although for a smaller set of courses and programs) and provide similar evidence.

[10] Notice that, given that we only use one cohort of students, we only observe one teacher-course observation and the process by which class identifiers change for the same professor over time is irrelevant for our analysis. We present it here only to provide a complete and clear picture of the institutional setting.

Table 3
Randomness checks – students.

Test statistics	Female	Academic high school[a]	High school grade	Entry test score	Top Income Bracket[a]	Outside Milan	Late enrollees[a]
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
	χ^2	χ^2	F	F	χ^2	χ^2	χ^2
Management							
Mean	0.489	0.482	0.497	0.393	0.500	0.311	0.642
Median	0.466	0.483	0.559	0.290	0.512	0.241	0.702
Minimum	0.049	0.055	0.012	0.004	0.037	0.000	0.025
Maximum	0.994	0.949	0.991	0.944	0.947	0.824	0.970
p-Value[b] (total number of tests is 20)							
<0.01	0	0	0	1	0	3	0
<0.05	1	0	1	1	2	6	1
Economics							
Mean	0.376	0.662	0.323	0.499	0.634	0.632	0.846
Median	0.292	0.715	0.241	0.601	0.616	0.643	0.911
Minimum	0.006	0.077	0.000	0.011	0.280	0.228	0.355
Maximum	0.950	0.993	0.918	0.989	0.989	0.944	0.991
p-Value[b] (total number of tests is 11)							
<0.01	1	0	2	0	0	0	0
<0.05	1	0	2	1	0	0	0
Law&Management							
Mean	0.321	0.507	0.636	0.570	0.545	0.566	0.948
Median	0.234	0.341	0.730	0.631	0.586	0.533	0.948
Minimum	0.022	0.168	0.145	0.182	0.291	0.138	0.935
Maximum	0.972	0.966	0.977	0.847	0.999	0.880	0.961
p-Value[b] (total number of tests is 7)							
<0.01	0	0	0	0	0	0	0
<0.05	2	0	0	0	0	0	0

The reported statistics are derived from probit (columns 1, 2, and 5–7) or OLS (columns 3 and 4) regressions of the observable students' characteristics (by column) on class dummies for each course in each degree program that we consider (Management: 20 courses, 144 classes; Economics: 11 courses, 72 classes; Law&Management: 7 courses, 14 classes). The reported p-values refer to tests of the null hypothesis that the coefficients on all the class dummies in each model are all jointly equal to zero. The test statistics are either χ^2 (columns 1, 2, and 5–7) or F (columns 3 and 4), with varying parameters depending on the model.

[a] See notes to Table 1.

[b] Number of courses for which the p-value of the test of joint significance of the class dummies is below 0.05 or 0.01.

they are located. There is no variation in other features of the rooms, such as the furniture (all rooms were fitted with exactly the same equipment: projector, computer, whiteboard) or the orientation (all rooms face the inner part of the campus where there is very limited car traffic).[11]

Table 4 provides evidence of the lack of correlation between teachers' and classes' characteristics, showing the results of regressions of teachers' observable characteristics on classes' observable characteristics. For this purpose, we estimate a system of 9 seemingly unrelated simultaneous equations, where each observation is a class in a compulsory course. The dependent variables are 9 teachers' characteristics (age, gender, h-index, average citations per year and 4 dummies for academic positions) and the regressors are the class characteristics listed in the rows of the table.[12] The reported statistics test the null hypothesis that the coefficients on each class characteristic are all jointly equal to zero in all the equations of the system.[13]

Results show that only the time of the lectures is significantly correlated with the teachers' observables at conventional statistical levels. In fact, this is one of the few elements of the teaching planning over which teachers had some limited choice. More specifically, professors are given a suggested time schedule for their classes and they can either approve it or request changes. The administration, then, accommodates such changes only if they are compatible with the other many constraints in terms of rooms availability and course overlappings. In our empirical analysis we do control for all the factors in Table 4, so that our measures of teaching effectiveness are purged from the potential confounding effect of teaching times on students' learning.

3. Estimating teacher effectiveness

We use student performance to estimate measures of teacher effectiveness. Namely, for each compulsory course

[11] In principle we could also condition on room fixed effects but there are several rooms in which only one class of the courses that we consider was taught.

[12] The h-index is a quality-adjusted measure of individual citations based on search results on Google Scholar. It was proposed by Hirsch (2005) and it is defined as follows: *A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each.*

[13] To construct the tests we use the small sample estimate of the variance–covariance matrix of the system.

we compare the future outcomes of students that attended those courses in different classes, under the assumption that students who were taught by better professors enjoyed better outcomes later on. This approach is similar to the *value-added* methodology commonly used in primary and secondary schools (Goldhaber & Hansen, 2010; Hanushek, 1979; Hanushek & Rivkin, 2006, 2010; Rivkin et al., 2005; Rothstein, 2009) but it departs from its standard version, that uses contemporaneous outcomes and conditions on past performance, since we use future performance to infer current teaching quality.[14]

One most obvious concern with the estimation of teacher quality is the non-random assignment of students to professors. For example, if the best students self-select themselves into the classes of the best teachers, then estimates of teacher quality would be biased upward. Rothstein (2009) shows that such a bias can be substantial even in well-specified models and especially when selection is mostly driven by unobservabes.

We avoid these complications by exploiting the random allocation of students in our cohort to different classes for each of their compulsory courses. We focus exclusively on compulsory courses, as self-selection is an obvious concern for electives.

We compute our measures of teacher effectiveness in two steps. First, we estimate the conditional mean of the future grades of students in each class according to the following procedure. Consider a set of students enrolled in degree program d and indexed by $i = 1, \ldots, N_d$, where N_d is the total number of students in the program. We have three degree programs ($d = \{1, 2, 3\}$): Management, Economics and Law&Management. Each student i attends a fixed sequence of compulsory courses indexed by $c = 1, \ldots, C_d$, where C_d is the total number of such compulsory courses in degree program d. In each course c the student is randomly allocated to a class $s = 1, \ldots, S_c$, where S_c is the total number of classes in course c. Denote by $\zeta \in Z_c$ a generic (compulsory) course, different from c, which student i attends in semester $t \geq t_c$, where t_c is the semester in which course c is taught. Z_c is the set of compulsory courses taught in any term $t \geq t_c$.

Let $y_{ids\zeta}$ denote the grade obtained by student i in course ζ. To control for differences in the distribution of grades across courses and to facilitate the interpretation of the results, $y_{ids\zeta}$ is standardized at the course level. Then, for each course c in each program d we run the following regression:

$$y_{ids\zeta} = \alpha_{dcs} + \beta X_i + \epsilon_{ids\zeta} \tag{1}$$

where X_i is a vector of student-level characteristics including a gender dummy, a dummy for whether the student is in the top income bracket, the entry test score and the high school leaving grade. The αs are our parameters of interest and they measure the conditional means of the future grades of students in class s: high values of α indicate that, on average, students attending course c in class s performed better (in subsequent courses) than students taking course c in a different class. The random allocation procedure guarantees that the class fixed effects α_{dcs} in Eq. (1) are purely exogenous and identification is straightforward.[15]

Eq. (1) does not include fixed effects for the classes in which students take each of the ζ courses since the random allocation guarantees that they are orthogonal to the αs, which are our main object of interest. Introducing such fixed effects would reduce the efficiency of the estimated αs without affecting their consistency.

Notice that, since in general there are several subsequent courses ζ for each course c, each student is observed multiple times and the error terms $\epsilon_{ids\zeta}$ are serially correlated within i and across ζ. We address this issue by adopting a standard random effect model to estimate all the equations (1), and we further allow for cross-sectional correlation among the error terms of students in the same class by clustering the standard errors at the class level.

More formally, we assume that the error term is composed of three additive components (all with mean equal zero):

$$\epsilon_{ids\zeta} = \nu_i + \omega_s + \nu_{ids\zeta} \tag{2}$$

where ν_i and ω_s are, respectively, an individual and a class component, and $\nu_{ids\zeta}$ is a purely random term. Operatively, we first apply the standard random effect transformation to the original model of Eq. (1).[16]

In the absence of other sources of serial correlation (i.e. if the variance of ω_s were zero), such a transformation would lead to a serially uncorrelated and homoskedastic variance–covariance matrix of the error terms, so that the standard random effect estimator could be produced by running simple OLS on the transformed model. In our specific case, we further cluster the transformed errors at the class level to account for the additional serial correlation induced by the term ω_s.

Overall, we are able to estimate 230 such fixed effects, the large majority of which are for Management courses.[17] Although this is admittedly not a particularly large

[14] For this reason we prefer to use the label *teacher effectiveness* for our estimates.

[15] Notice that in few cases more than one teacher taught in the same class, so that our class effects capture the overall effectiveness of teaching and cannot be attached to a specific person. Since the students' evaluations are also available at the class level and not for specific teachers, we cannot disaggregate further.

[16] The standard random effect transformation subtracts from each variable in the model (both the dependent and each of the regressors) its within-mean scaled by the factor $\theta = 1 - \sqrt{\sigma_\nu^2/(|Z_c|(\sigma_\omega^2 + \sigma_\nu^2) + \sigma_\nu^2}$), where $|Z_c|$ is the cardinality of Z_c. For example, the random-effects transformed dependent variable is $y_{ids\zeta} - \theta \overline{y}_{ids}$, where $\overline{y}_{ids} = |Z_c|^{-1} \sum_{h=1}^{|Z_c|} y_{idh\zeta}$. Similarly for all the regressors. The estimates of σ_ν^2 and $(\sigma_\omega^2 + \sigma_\nu^2)$ that we use for this transformation are the usual Swamy–Arora, also used by the command *xtreg* in Stata (Swamy & Arora, 1972).

[17] We cannot run Eq. (1) for courses that have no contemporaneous nor subsequent courses, such as Corporate Strategy for Management, Banking for Economics and Business Law for Law&Management (see Table A.1). For such courses, the set Z_c is empty. Additionally, some courses in Economics and in Law&Management are taught in one single class, for example Econometrics (for Economics students) or Statistics (for Law&Management). For such courses, we have $S_c = 1$. The evidence that we reported in Tables 3 and 4 also refer to the same set of 230 classes.

Table 4
Randomness checks – teachers.

	F-test	*p*-Value
Class size[a]	0.94	0.491
Attendance[b]	0.95	0.484
Avg. high school grade	0.73	0.678
Avg. entry test score	1.37	0.197
Share of females	1.05	0.398
Share of students from outside Milan[c]	0.25	0.987
Share of top-income students[c]	1.31	0.228
Share academic high school[c]	1.35	0.206
Share late enrollees[c]	0.82	0.597
Share of high ability[d]	0.69	0.716
Share of early morning lectures	5.24	0.000
Share of late afternoon lectures	1.97	0.039
Room's floor[e]	0.45	0.998
Dummy for building A	1.39	0.188

The reported statistics are derived from a system of 9 seemingly unrelated simultaneous equations, where each observation is a class in a compulsory course (184 observations in total). The dependent variables are 9 teachers' characteristics (age, gender, *h*-index, average citations per year and 4 dummies for academic positions) and the regressors are the class characteristics listed in the table. The reported statistics test the null hypothesis that the coefficients on each class characteristic are all jointly equal to zero in all the equations of the system. The last row tests the hypothesis that the coefficients on all regressors are all jointly zero in all equations. All tests are distributed according to a *F*-distribution with (9,1467) degrees of freedom, apart from the joint test in the last row, which has (108,1467) degrees of freedom.

[a] Number or officially enrolled students.

[b] Attendance is monitored by random visits of university attendants to the class.

[c] See notes to Table 1.

[d] Share of students in the top 25% of the entry test score distribution.

[e] Test of the joint significance of 4 floor dummies.

number, it compares favorably with other studies in the literature. For example, Carrell and West (2010) only observe 91 professors, while Krautmann and Sander (1999) and Weinberg et al. (2009) have 258 and 395 observations, respectively.

The second step of our approach is meant to purge the estimated $\hat{\alpha}$ from the effect of other class characteristics that might affect the performance of students in later courses but are not attributable to teachers. By definition, the class fixed effects capture all features, both observable and unobservable, that are fixed for all students in the class. These certainly include teaching quality but also other factors that are documented to be important ingredients of the education production function, such as class size and composition (De Giorgi, Pellizzari, & Woolston, 2012).

A key advantage of our data is that most of these other factors are observable. Based on our academic records we can construct measures of both class size and class composition (in terms of students' characteristics). Additionally, we also have access to the identifiers of the teachers in each class and we can recover a large set of variables like gender, tenure status, and measures of research output. We also know which of the several teachers in each course acted as coordinator. These are the same teacher characteristics that we used in Table 4. Once we condition on all these observable controls, unobservable teaching quality is likely to be the only remaining factor that generates variation in the estimated $\hat{\alpha}$. At a minimum, it should be uncontroversial that teaching quality is by far the single most important unobservable that generates variation in the $\hat{\alpha}$s, once conditioning on the observables.

The effect of social interactions among the students might also affect the estimated $\hat{\alpha}$s. However, notice that if such effects are related to the observable characteristics of the students, then we are able to control for those. Additionally, there might be complementarities among teachers' ability and students' interactions, as good teachers are also those who stimulate fruitful collaborations among their students. This component of the social interaction effects is certainly something that one would like to incorporate in a measure of teaching quality, as in our analysis.

Thus, in Table 5 we regress the estimated $\hat{\alpha}$ on all observable class and teacher characteristics. In column 1 we condition only on class size and class composition, in column 2 only on information about the teachers and in column 3 we combine the two sets of controls. In all cases we weight observations by the inverse of the standard error of the estimated $\hat{\alpha}$ to take into account differences in the precision of such estimates. Consistently with previous studies on the same data (De Giorgi et al., 2012), we find that larger classes tend to be associated with worse learning outcomes, that classes with more able students, measured with either high school grades or the entry test score, also perform better and that a high concentration of high income students appears to be detrimental for learning. Overall, observable class characteristics explain about 8% of the variation in the estimated $\hat{\alpha}$ within degree program, term and subject cells, where subjects are defined as in Table A.1.[18]

The results in column 2 show a non-linear relationship between teachers' age and teaching outcomes, which might be rationalized with increasing returns to experience. Also, professors who are more productive in research seem to be less effective as teachers, when output is measured with the *h*-index. The effect is reversed using yearly citations but it never reaches acceptable levels of statistical significance. Finally, and consistently with the age effect, also the professor's academic position matters, with a ranking that gradually improves from assistant to associate to full professors (other academic positions, such as external or non tenured-track teachers, are the excluded group). However, as in Hanushek and Rivkin (2006) and Krueger (1999), we find that the individual traits of the teachers explain less than a tenth of the (residual) variation in students' achievement. Overall, the complete set of observable class and teachers' variables explains approximately 15% of the (residual) variation.

[18] The Partial *R*-squared reported at the bottom of the table refer to the *R*-squared of a partitioned regression where the dummies for the degree program, the term and the subject are partialled out. The total *R*-squared of the regressions in Table 5 are 0.812, 0.810 and 0.826 for column one, two and three, respectively.

Table 5
Determinants of class effects.

Dependent variable = $\hat{\alpha}_s$	[1]	[2]	[3]
Class size[a]	−0.000**	–	−0.000**
	(0.000)		(0.000)
Avg. HS grade	2.159**	–	2.360**
	(1.064)		(1.070)
Avg. entry test score	−1.140	–	−1.530
	(1.426)		(1.405)
Share of females	0.006	–	−0.094
	(0.242)		(0.245)
Share from outside Milan	−0.080	–	−0.078
	(0.208)		(0.201)
Share of top income[a]	−0.283	–	−0.331
	(0.277)		(0.278)
Share from academic HS	0.059	–	−0.054
	(0.308)		(0.313)
Share of late enrollees	−0.365	–	0.017
	(0.847)		(0.843)
Share of high ability[a]	0.733*	–	0.763*
	(0.404)		(0.390)
Morning lectures[a]	0.015	–	−0.015
	(0.038)		(0.040)
Evening lectures[a]	−0.175	–	−0.170
	(0.463)		(0.490)
1 = coordinator	–	0.013	0.039
		(0.038)	(0.041)
Male	–	−0.017	−0.014
		(0.024)	(0.025)
Age	–	−0.013***	−0.013**
		(0.005)	(0.005)
Age squared	–	0.000**	0.000*
		(0.000)	(0.000)
H-index	–	−0.008	−0.007
		(0.006)	(0.006)
Citations per year	–	0.000	0.000
		(0.001)	(0.001)
Full professor		0.116*	0.121*
		(0.066)	(0.072)
Associate professor		0.113*	0.118*
		(0.062)	(0.067)
Assistant professor		0.109*	0.123*
		(0.061)	(0.065)
Classroom characteristics[b]	Yes	No	Yes
Degree program dummies	Yes	Yes	Yes
Subject area dummies	Yes	Yes	Yes
Term dummies	Yes	Yes	Yes
Partial *R*-squared[c]	0.089	0.081	0.158
Observations	230	230	230

Observations are weighted by the inverse of the standard error of the estimated αs. In column 3, variables are weighted averages of individual characteristics if there is more than one teacher per class. All variables regarding the academic position refer to the main teacher of the class. The excluded dummy is a residual category (visiting prof., external experts, collaborators).

[a] See notes to Table 4.

[b] Four floor dummies, one building dummy and a dummy for multi-classrooms classes.

[c] *R*-squared computed once program, term and subject fixed effects are partialled out.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Our final measures of teacher effectiveness are the residuals of the regression of the estimated $\hat{\alpha}$ on all the observable variables, i.e. the regression reported in column 3 of Table 5. Such residuals are denoted τ_{cds}.[19]

Finally, in order to estimate the variance of estimated teacher effectiveness across classes within courses, we have to take into account the variance of the estimation error. Following Weinberg et al. (2009), we randomly split in half all classes in our sample and we replicate our estimation procedure, obtaining for each class two estimates of τ_{cds}, denoted τ' and τ''.[20] The random split ensures that the estimation errors in τ' and τ'' are orthogonal to each other, so that the variance of τ can be estimated as the covariance between τ' and τ''.[21] In Table 6 we present descriptive statistics of such measures. Given the relatively small size of our sample of class effects (230), the variance due to the estimation error is relatively large and the correction procedure described above is particularly important. About 60% of the unadjusted variance of the teacher effects is accounted for by the estimation error.

The overall standard deviation of teacher effectiveness is 0.038. This average is the composition of a larger variation among the courses of the program in Economics (0.086) and a more limited variation in Management (0.021) and Law&Management (0.012). Grades are normalized, so that the distributions of the class effects are comparable across courses and these results can be directly interpreted in terms of changes in outcomes. The overall effect of increasing teacher effectiveness by one standard deviation is an increase in the average grade of subsequent courses by 0.038 standard deviations, roughly 0.1 of a grade point or 0.5% over the average grade of approximately 26.[22] For comparison, Carrell and West (2010) estimate that a one-standard deviation increase of teacher quality would increase the achievement of the students at their military academy by about 0.05 of a standard deviation, a result that is slightly larger but very comparable to our estimated average effect. The non-

[19] Theoretically, the choice to purge the first-stage residuals of both the class and the teacher's observable characteristics could be controversial. If this exercise aims at measuring teacher effectiveness one may prefer not to condition on the teacher's individual characteristics, insofar as they can themselves be correlated with the unobservable trait that we want to measure. In practice, however, the results obtained conditioning or not on teacher's observables are extremely similar due to the fact that such observables explain a very small fraction of the variation in students' grades. Results that use the residuals of the regression in column 2 of Table 5 (as opposed to column 3) as a measure of teacher effectiveness are available from the authors upon request.

[20] This procedure is replicated 50 times.

[21] To be more precise, this holds true assuming that any effect of social interactions is captured by either observable students' characteristics or by the class effect.

[22] In Italy, university exams are graded on a scale 0–30, with pass equal to 18. Such a peculiar grading scale comes from historical legacy: while in primary, middle and high school students were graded by one teacher per subject on a scale 0–10 (pass equal to 6), at university each exam was supposed to be evaluated by a commission of three professors, each grading on the same 0–10 scale, the final mark being the sum of these three. Hence, 18 is pass and 30 is full marks. Apart from the scaling, the actual grading at Bocconi is performed as in the average US or UK university.

Table 6
Descriptive statistics of estimated teacher effectiveness.

	Management	Economics	Law&Management	Total
Panel A: standard deviation of teacher effectiveness				
Mean	0.021	0.086	0.012	0.038
Minimum	0.003	0.015	0.000	0.000
Maximum	0.040	0.140	0.035	0.140
Panel B: largest minus smallest teacher effectiveness				
Mean	0.190	0.432	0.027	0.230
Minimum	0.123	0.042	0.014	0.014
Maximum	0.287	0.793	0.043	0.793
No. of courses	20	11	7	38
No. of classes	144	72	14	230

Teacher effectiveness is estimated by regressing the estimated class effects (α) on observable class and teacher's characteristics (see Table 5). Standard deviation is computed as the covariance between estimated teacher effects for randomly selected subgroups of the class.

random assignment of students to teachers in Weinberg et al. (2009) makes it more difficult to compare their estimates, which are in the order of 0.15–0.2 of a standard deviation, with ours.

To put the magnitude of these estimates into perspective, it is useful to compare them with the effects of other inputs of the education production function that have been estimated in the literature. For example, the many papers that have investigated the effect of changing the size of the classes on students' academic performance (Angrist & Lavy, 1999; Bandiera, Larcinese, & Rasul, 2010; Krueger, 1999) present estimates in the order of 0.1–0.15 of a standard deviation for a 1-standard deviation change of class size. Our results suggest that the effect of teachers is approximately 25–40% of that of class size.[23]

In Table 6 we also report the standard deviations of teacher effectiveness of the courses with the least and the most variation to show that there is substantial heterogeneity across courses. Overall, we find that in the course with the highest variation (Macroeconomics in the Economics program) the standard deviation of our measure of effectiveness is approximately 15% of a standard deviation in grades (almost 2% of the average grade). This compares to a standard deviation of essentially zero in the courses with the lowest variation (Mathematics and Accounting in the Law&Management program).

In the lower panel of Table 6 we show the mean (across courses) of the difference between the largest and the smallest indicators of teacher effectiveness, which allows us to compute the effect of attending a course in the class of the best versus the worst teacher. On average, this effect amounts to 0.230 of a standard deviation, that is almost 0.8 grade points or 3% over the average grade. This average effect masks a large degree of heterogeneity across subjects ranging from almost 80% to a mere 4% of a standard deviation.

To further understand the importance of these effects, we can also compare particularly lucky students, who are assigned to good teachers (those in the top 5% of the distribution of effectiveness) throughout their sequence of compulsory courses, to particularly unlucky students, who are always assigned to bad teachers (those in the bottom 5% of the distribution of effectiveness). The average grades of these two groups of students are 1.8 grade points apart, corresponding to over 7% of the average grade.

For robustness and comparison, we estimate the class effects in two alternative ways. First, we restrict the set Z_c to courses belonging to the same subject area of course c, under the assumption that good teaching in one course has a stronger effect on learning in courses of the same subject areas (e.g. a good basic mathematics teacher is more effective in improving students performance in econometrics than in business law). The subject areas are defined by the colors in Table A.1 and correspond to the department that was responsible for the organization and teaching of the course. We label these estimates *subject* effects. Given the more restrictive definition of Z_c we can only produce these estimates for a smaller set of courses and using fewer observation, which is why we do not take them as our benchmark.

Next, rather than using performance in subsequent courses, we run Eq. (1) with the grade in the same course c as the dependent variable. We label these estimates *contemporaneous* effects. We do not consider these contemporaneous effects as alternative and equivalent measures of teacher effectiveness, but we will use them to show that they correlate very differently with the students' evaluations.

In order to investigate the correlation between these alternative estimates of teacher effectiveness in Table 7 we

Table 7
Comparison of benchmark, subject and contemporaneous teacher effects.

Dependent variable: benchmark teacher effectiveness		
Subject	0.048** (0.023)	–
Contemp.	–	−0.096*** (0.019)
Program fixed effects	Yes	Yes
Term fixed effects	Yes	Yes
Subject fixed effects	Yes	Yes
Observations	212	230

Bootstrapped standard errors in parentheses. Observations are weighted by the inverse of the standard error of the dependent variable.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

[23] Notice, however, that the size of the class influences students' performance through the behavior of the teachers, at least partly.

Table 8
Teacher effectiveness and students' evaluations.

	Teaching quality		Lecturing clarity		Teacher ability in generating interest		Course logistics		Course workload	
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Teacher effectiveness										
Benchmark	−0.496**	–	−0.249**	–	−0.552**	–	−0.124	–	−0.090	–
	(0.236)		(0.113)		(0.226)		(0.095)		(0.104)	
Contemporaneous	–	0.238***	–	0.116***	–	0.214***	–	0.078***	–	−0.007
		(0.055)		(0.029)		(0.044)		(0.019)		(0.025)
Class characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher's characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Degree program dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject area dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Term dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Partial R^2	0.019	0.078	0.020	0.075	0.037	0.098	0.013	0.087	0.006	0.001
Observations	230	230	230	230	230	230	230	230	230	230

Weighted OLS estimates. Observations are weighted by the number of collected questionnaires in each class. Bootstrapped standard errors in parentheses.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

run two weighted OLS regressions with our benchmark estimates as the dependent variable and, in turn, the subject and the contemporaneous effects on the right hand side, together with dummies for degree program, term and subject area.

Reassuringly, the subject effects are positively and significantly correlated with our benchmark, while the contemporaneous effects are negatively and significantly correlated with our benchmark, a result that is consistent with previous findings (Carrell & West, 2010; Krautmann & Sander, 1999; Weinberg et al., 2009).

4. Correlating teacher effectiveness and student evaluations

In this section we investigate the relationship between our measures of teaching effectiveness and the evaluations teachers receive from their students. We concentrate on two core items from the evaluation questionnaires, namely overall teaching quality and the overall clarity of the lectures. Additionally, we also look at other items: the teacher's ability in generating interest for the subject, the logistics of the course (schedule of classes, combinations of practical sessions and traditional lectures) and the total workload compared to other courses.

Formally, we estimate the following equation:

$$q^k_{dtcs} = \lambda_0 + \lambda_1 \hat{\alpha}_{dtcs} + \lambda_2 C_{dtcs} + \lambda_3 T_{dtcs} + \gamma_d + \delta_t + \upsilon_c + \epsilon_{dtcs} \quad (3)$$

where q^k_{dtcs} is the average answer to question k in class s of course c in degree program d (which is taught in term t), $\hat{\alpha}_{dtcs}$ is the estimated class fixed effect from Eq. (1), C_{dtcs} is the set of class characteristics, T_{dtcs} is the set of teacher characteristics. γ_d, δ_t and υ_c are fixed effects for degree program, term and subject areas, respectively. ϵ_{dtcs} is a residual error term.

Notice that the class and teacher characteristics are exactly the same as in Table 5, so that Eq. (3) is equivalent to a partitioned regression model of the evaluations q_{dtcs} on our measures of teacher effectiveness, i.e. the residuals of the regressions in Table 5, where all the observables and the fixed effects are partialled out.

Since the dependent variable in Eq. (3) is an average, we use weighted OLS, where each observation is weighted by the square root of the number of collected questionnaires in the class, which corresponds to the size of the sample over which the average answers are taken. Additionally, we also bootstrap the standard errors to take into account the presence of generated regressors (the $\hat{\alpha}$s).

Table 8 reports the estimates of Eq. (3) for single evaluation items. For each item we show results obtained using our benchmark estimates of teacher effectiveness and those obtained using the contemporaneous class effects. For convenience, results are reported graphically in Fig. 1.

Our benchmark class effects are negatively associated with all the items that we consider, suggesting that teachers who are more effective in promoting future performance receive worse evaluations from their students. This relationship is statistically significant for all items (but logistics), and is of sizable magnitude. For example, a one-standard deviation increase in teacher effectiveness reduces the students' evaluations of overall teaching quality by about 50% of a standard deviation. Such an effect could move a teacher who would otherwise receive a median evaluation down to the 31st percentile of the distribution. Effects of slightly smaller magnitude can be computed for lecturing clarity.

When we use the contemporaneous effects the estimated coefficients turn positive and highly significant for all items (but workload). In other words, the teachers of classes that are associated with higher grades in their own exam receive better evaluations from their students. The magnitudes of these effects is smaller than those estimated for our benchmark measures: one standard deviation change in the contemporaneous teacher effect increases the evaluation of overall teaching quality by 24% of a standard deviation and the evaluation of lecturing clarity by 11%.

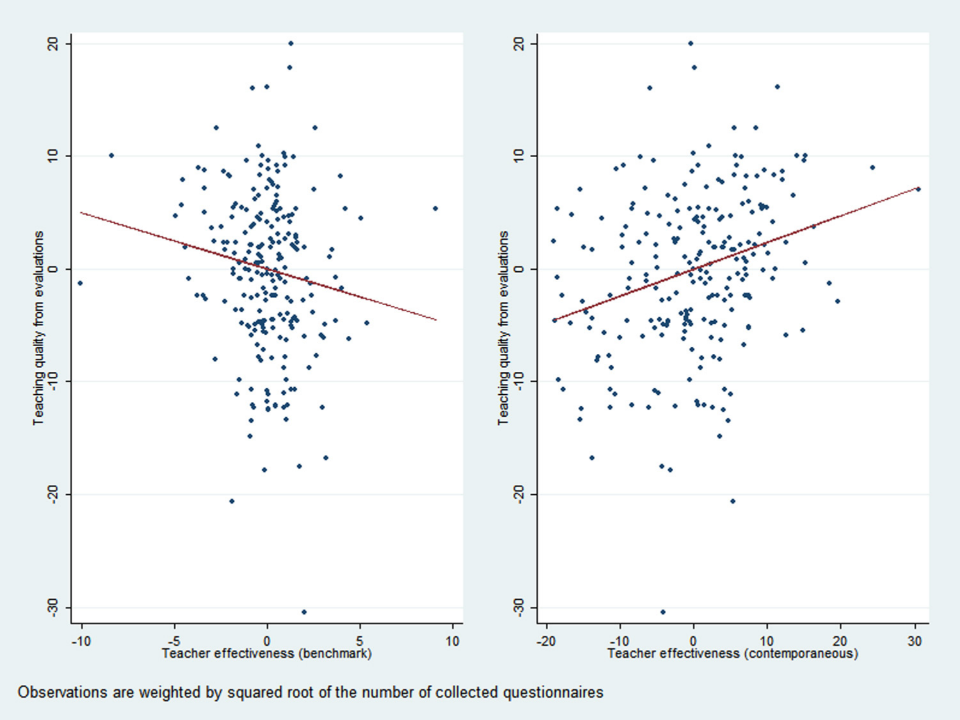

Fig. 1. Students' evaluations and estimated teacher effectiveness.

These results are broadly consistent with the findings of other studies (Carrell & West, 2010; Krautmann & Sander, 1999; Weinberg et al., 2009). Krautmann and Sander (1999) only look at the correlation of students evaluations with (expected) contemporaneous grades and find that a one-standard deviation increase in classroom GPA results in an increase in the evaluation of between 0.16 and 0.28 of a standard deviation in the evaluation. Weinberg et al. (2009) estimate the correlation of the students' evaluations of professors and both their current and future grades and report that "a one standard deviation change in the current course grade is associated with a large increase in evaluations-more than a quarter of the standard deviation in evaluations", a finding that is very much in line with our results. When looking at the correlation with future grades, Weinberg et al. (2009) do not find significant results.[24] The comparison with the findings in Carrell and West (2010) is complicated by the fact that in their analysis the dependent variable is the teacher effect whereas the items of the evaluation questionnaires are on the right-hand-side of their model. Despite these difficulties, the positive correlation of evaluations and current grades and the reversal to negative correlation with future grades is a rather robust finding.

These results clearly challenge the validity of students' evaluations of professors as a measure of teaching quality. Even abstracting from the possibility that professors strategically adjust their grades to please the students (a practice that is made difficult by the timing of the evaluations, that are always collected before the exam takes place), it might still be possible that professors who make the classroom experience more enjoyable do that at the expense of true learning or fail to encourage students to exert effort. Alternatively, students might reward teachers who prepare them for the exam, that is teachers who teach to the test, even if this is done at the expenses of true learning. This interpretation is consistent with the results in Weinberg et al. (2009), who provide evidence that students are generally unaware of the value of the material they have learned in a course.

0.1pt?>Of course, one may also argue that students' satisfaction is important per se and, even, that universities should aim at maximizing satisfaction rather than learning, especially private institutions like Bocconi. We doubt that this is the most common understanding of higher education policy.

5. Robustness checks

In this section we present some robustness checks for our main results.

First, one might be worried that students might not comply with the random assignment to the classes. For various reasons they may decide to attend one or more courses in a different class from the one to which they were formally allocated.[25] Unfortunately, such changes would

[24] Notice also that Weinberg et al. (2009) consider average evaluations taken over the entire career of professors.

[25] For example, they may wish to stay with their friends, who might have been assigned to a different class, or they may like a specific teacher, who is known to present the subject particularly clearly.

Table 9
Robustness check for class switching.

	Overall teaching quality		Lecturing clarity	
	[1]	[2]	[3]	[4]
Panel A: all courses				
Benchmark teacher effects	−0.496**	–	−0.249**	–
	(0.236)		(0.113)	
Contemporaneous teacher effects	–	0.238***	–	0.116***
		(0.055)		(0.029)
Observations	230	230	230	230
Panel B: excluding most switched course				
Benchmark teacher effects	−0.572**	–	−0.261**	–
	(0.267)		(0.118)	
Contemporaneous teacher effects	–	0.258***	–	0.121***
		(0.064)		(0.030)
Observations	222	222	222	222
Panel C: excluding most and second most switched course				
Benchmark teacher effects	−0.505*	–	−0.234*	–
	(0.272)		(0.128)	
Contemporaneous teacher effects	–	0.233***	–	0.112***
		(0.062)		(0.031)
Observations	214	214	214	214
Panel D: excluding five most switched courses				
Benchmark teacher effects	−0.579**	–	−0.229*	–
	(0.273)		(0.122)	
Contemporaneous teacher effects	–	0.154**	–	0.065**
		(0.063)		(0.032)
Observations	176	176	176	176

Weighted OLS estimates. Observations are weighted by the squared root of the number of collected questionnaires in each class. Additional regressors: teacher characteristics (gender and coordinator status), class characteristics (class size, attendance, average high school grade, average entry test score, share of high ability students, share of students from outside Milan, share of top-income students), degree program dummies, term dummies, subject area dummies. Bootstrapped standard errors in parentheses.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

not be recorded in our data, unless the student formally asked to be allocated to a different class, a request that needed to be adequately motivated.[26] Hence, we cannot exclude a priori that some students switch classes.

If the process of class switching is unrelated to teaching quality, then it merely affects the precision of our estimated class effects, but it is very well possible that students switch in search for good or lenient lecturers. We can get some indication of the extent of this problem from the students' answers to an item of the evaluation questionnaire that asks about the congestion in the classroom. Specifically, the question asks whether the number of students in the class was detrimental to one's learning. We can, thus, identify the most congested classes from the average answer to such question in each course.

Courses in which students concentrate in the class of one or few professors should be characterized by a very skewed distribution of such a measure of congestion, with one or a few classes being very congested and the others being pretty empty. Thus, for each course we compute the difference in the congestion indicator between the most and the least congested classes (over the standard deviation). Courses in which such a difference is very large should be the ones that are more affected by switching behaviors.

In Table 9 we replicate our benchmark estimates for two core evaluation items (overall teaching quality and lecturing clarity) by excluding (in Panel B) the most switched course, i.e. the course with the largest difference between the most and the least congested classes (which is marketing). For comparison, we also report the original estimates from Table 8 in Panel A and we find that results change only marginally. Next, in Panels C and D we exclude from the sample also the second most switched course (human resource management) and the five most switched courses, respectively.[27] Again, the estimated coefficients are only mildly affected, although the significance levels are reduced according with the smaller sample sizes.

[26] Possible motivations for such requests could be health reasons. For example, due to a broken leg a student might not be able to reach classrooms in the upper floors of the university buildings and could ask to be assigned to a class taught on the ground floor.

[27] The five most switched courses are marketing, human resource management, mathematics for Economics and Management, financial mathematics and managerial accounting.

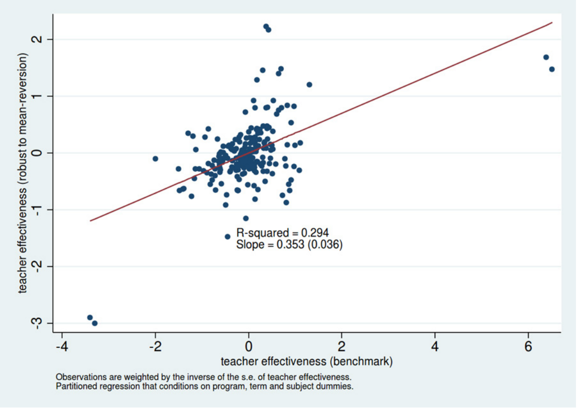

Fig. 2. Robustness check for mean reversion in grades.

Overall, course switching does not seem to affect our estimates in any major direction.

Another possible concern is that our results may be generated by some endogenous reaction of students to the quality of their past teachers. For example, meeting a bad teacher might induce exerting higher effort in the future to catch up, especially if bad teaching resulted in a lower (contemporaneous) grade. Hence, the students evaluations may reflect real teaching quality and our measure of teacher effectiveness would be biased by such a process of mean reversion, leading to a negative correlation with real teaching quality and, consequently, also with the evaluations of the students.

To control for this potential feedback effect on students' effort we recompute our benchmark measures of teacher effectiveness adding the student average grade in all previous courses to the set of controls. Fig. 2 shows that these two measures are strongly correlated.

6. Interpretation and further evidence

The interpretation of the students' evaluations as measures of the quality of teaching rests on the – explicit or implicit – view that the students observe the quality of teaching in the classroom and, when asked to report it in the questionnaire, they do so truthfully. Our results, however, contradict this view and seem more consistent with the idea that students evaluate teachers on the basis of their enjoyment of the course or, in the words of economists, on the basis of their realized utility. Good teachers – those who provide their students with knowledge that is useful in future learning – presumably require their students to exert effort by paying attention and being concentrated in class and by doing demanding homework. As it is commonly assumed in economic models, agents dislike exerting effort and, if the students' questionnaires reflect utility, it is very possible that good teachers are badly evaluated by their students.

To provide further support for this interpretation of our results, in this section we present two additional pieces of evidence.

First, in order to support the claim that the students' questionnaires reflect the students' enjoyment of the class experience rather than the quality of teaching, Table 10 shows that the evaluations are significantly affected by weather conditions on the day in which they were filled. There is ample evidence that people's utility (or welfare, happiness, satisfaction) improves with good meteorological conditions (Barrington-Leigh, 2008; Connolly, 2013; Denissen, Butalid, Penke, & van Aken, 2008; Keller et al., 2005; Schwarz & Clore, 1983) and finding that such

Table 10
Students' evaluations and weather conditions.

	Teaching quality		Lecturing clarity		Teacher ability in generating interest		Course logistics		Course workload	
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Av. temperature	0.139*	0.120	0.063*	0.054	0.171***	0.146***	0.051**	0.047**	0.057*	0.053*
	(0.074)	(0.084)	(0.036)	(0.038)	(0.059)	(0.054)	(0.020)	(0.019)	(0.031)	(0.029)
1 = rain	−0.882**	−0.929**	−0.293	−0.314	−0.653**	−0.716**	−0.338***	−0.348***	0.081	0.071
	(0.437)	(0.417)	(0.236)	(0.215)	(0.327)	(0.287)	(0.104)	(0.108)	(0.109)	(0.128)
1 = fog	0.741**	0.687*	0.391**	0.367**	0.008	−0.063	0.303***	0.292***	−0.254***	−0.265***
	(0.373)	(0.377)	(0.191)	(0.170)	(0.251)	(0.247)	(0.085)	(0.090)	(0.095)	(0.096)
Teaching effectiveness	–	−0.424*	–	−0.189	–	−0.566**	–	−0.090	–	−0.088
		(0.244)		(0.120)		(0.223)		(0.088)		(0.093)
Class characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher's characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Degree program dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject area dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Term dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	230	230	230	230	230	230	230	230	230	230

Weighted OLS estimates. Observations are weighted by the number of collected questionnaires in each class. Bootstrapped standard errors in parentheses.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Table 11
Teacher effectiveness and students evaluations by share of high ability students.

	Presence of high-ability students			
	All	>0.22 (top 75%)	>0.25 (top 50%)	>0.27 (top 25%)
	[1]	[2]	[3]	[4]
Panel A: overall teaching quality				
Teaching effectiveness	-0.496^{**}	-0.502^{*}	-0.543	-0.141^{***}
	(0.236)	(0.310)	(0.439)	(0.000)
Panel B: lecturing clarity				
Teaching effectiveness	-0.249^{**}	-0.240	-0.283	-0.116^{*}
	(0.113)	(0.140)	(0.191)	(0.068)
Observations	230	171	114	56

Weighted OLS estimates. Observations are weighted by the squared root of the number of collected questionnaires in each class. Additional regressors: teacher characteristics (gender and coordinator status), class characteristics (class size, attendance, average high school grade, average entry test score, share of high ability students, share of students from outside Milan, share of top-income students), degree program dummies, term dummies, subject area dummies. Bootstrapped standard errors in parentheses.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

conditions also affect the evaluations of professors suggests that they indeed reflect utility rather than (or together with) teaching quality.

Specifically, we find that evaluations improve with temperature and in foggy days, and deteriorate in rainy days. The effects are significant for most of the items that we consider and the signs of the estimates are consistent across items and specifications.

Obviously, teachers might be affected by meteorological conditions as much as their students and one may wonder whether the estimated effects in the odd columns of Table 10 reflect the indirect effect of the weather on teaching effectiveness. We consider this interpretation to be very unlikely since the questionnaires are distributed and filled before the lecture. Nevertheless, we also condition on our benchmark measure of teaching effectiveness and, as we expected, we find that the estimated effects of both the weather conditions and teacher effectiveness itself change only marginally.

Second, if the students who dislike exerting effort are the least able, as it is assumed for example in the signaling model of schooling (Spence, 1973), we expect the correlation between our measures of teacher effectiveness and the average students' evaluations to be less negative in classes where the share of high ability students is higher. We define as high ability those students who score in the upper quartile of the distribution of the entry test score and, for each class in our data, we compute the share of such students. Then, we investigate the relationship between the students' evaluations and teacher effectiveness by restricting the sample to classes in which high-ability students are over-represented. Results shown in Table 11 seem to suggest the presence of non-linearities or threshold effects, as the estimated coefficient remains relatively stable until the fraction of high ability students in the class goes above one quarter or, more precisely, 27% which corresponds to the top 25% of the distribution of the presence of high ability students. At that point, the estimated effect of teacher effectiveness on students' evaluations is about a quarter of the one estimated on the entire sample. The results, thus, suggest that the negative correlations reported in Table 8 are mostly due to classes with a particularly low incidence of high ability students.

7. Policies and conclusions

Using administrative archives from Bocconi University and exploiting random variation in students' allocation to teachers within courses we find that, on average, students evaluate positively classes that give high grades and negatively classes that are associated with high grades in subsequent courses. These empirical findings challenge the idea that students observe the ability of the teacher in the classroom and report it to the administration when asked in the questionnaire. A more appropriate interpretation is based on the view that good teachers are those who require their students to exert effort; students dislike it, especially the least able ones, and their evaluations reflect the utility they enjoyed from the course.

Overall, our results cast serious doubts on the validity of students' evaluations of professors as measures of teaching quality or effort. At the same time, the strong effects of teaching quality on students' outcomes suggest that improving the quantity or the quality of professors' inputs in the education production function can lead to large gains. In the light of our findings, this could be achieved through various types of interventions.

First, since the evaluations of the best students are more aligned with actual teachers' effectiveness, the opinions of the very good students could be given more weight in the measurement of professors' performance. In order to do so, some degree of anonymity of the evaluations must be lost but there is no need for the teachers to identify the students: only the administration should be allowed to do it, and there are certainly ways to make the separation of information between administrators and professors credible to the students so as not to bias their evaluations. Second, one may think of adopting exam formats that reduce the returns to teaching-to-the-test, although this may come at larger costs due to the additional time needed to grade less standardized tests. At the same time, the extent of grade leniency could be greatly limited by making sure that teaching and grading are done by different persons.

Alternatively, questionnaires could be administered at a later point in the academic track to give students the time to appreciate the real value of teaching in subsequent learning (or even in the market). Obviously, this would also pose problems in terms of recall bias and possible retaliation for low grading.

Alternatively, one may also think of other forms of performance measurement that are more in line with the peer-review approach adopted in the evaluation of research output. It is already common practice in several departments to have colleagues sitting in some classes and observing teacher performance, especially of assistant professors. This is often done primarily with the aim of offering advise, but it could also be used to measure outcomes. To avoid teachers adapting their behavior due to the presence of the observer, teaching sessions could be recorded and a random selection of recordings could be evaluated by an external professor in the same field.

Obviously, these measurement methods – as well as other potential alternative are costly, but they should be compared with the costs of the current systems of collecting students' opinions about teachers, which are often non-trivial.

Appendix A

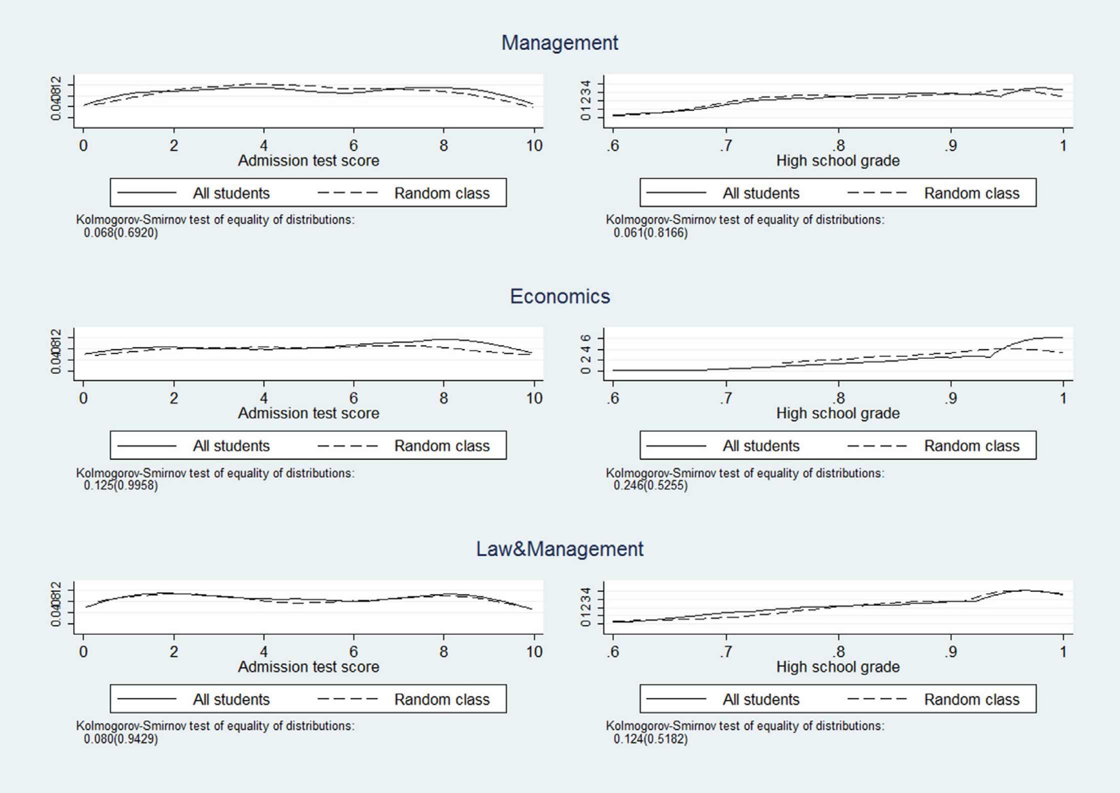

Fig. A.1. Evidence of random allocation – ability variables.

Table A.1
Structure of degree programs.

	MANAGEMENT	ECONOMICS	LAW&MANAG.
Term I	Management I	Management I	Management I
	Private law	Private law	Mathematics
	Mathematics	Mathematics	
Term II	Microeconomics	Microeconomics	Accounting
	Public law	Public law	
	Accounting	Accounting	
Term III	Management II	Management II	Management II
	Macroeconomics	Macroeconomics	Statistics
	Statistics	Statistics	
Term IV	Business law	Financial mathematics	Accounting II
	Manag. of Public Administrations	Public economics	Fiscal law
	Financial mathematics	Business law	Financial mathematics
	Human resources management		
Term V	Banking	Econometrics	Corporate finance
	Corporate finance	Economic policy	
	Management of industrial firms		
Term VI	Marketing	Banking	
	Management III		
	Economic policy		
	Managerial accounting		
Term VII	Corporate strategy		
Term VIII			Business law II

The colors indicate the subject area the courses belong to: red = management, black = economics, green = quantitative, and blue = law. Only compulsory courses are displayed.

Table A.2
Wording of the evaluation questions.

Overall teaching quality	On a scale 0–10, provide your overall evaluation of the course you attended in terms of quality of the teaching.
Clarity of the lectures	On a scale 1–5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the speech and the language of the teacher during the lectures are clear and easily understandable.
Ability in generating interest for the subject	On a scale 0–10, provide your overall evaluation about the teacher's ability in generating interest for the subject.
Logistics of the course	On a scale 1–5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the course has been carried out coherently with the objectives, the content and the schedule that were communicated to us at the beginning of the course by the teacher.
Workload of the course	On a scale 1–5, where 1 means complete disagreement and 5 complete agreement, indicate to what extent you agree with the following statement: the amount of study materials required for the preparation of the exam has been realistically adequate to the objective of learning and sitting the exams of all courses of the term.

References

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics, 114*, 533–575.

Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics, 109*, 1125–1156.

Bandiera, O., Larcinese, V., & Rasul, I. (2010). Heterogeneous class size effects: New evidence from a panel of university students. *Economic Journal, 120*, 1365–1398.

Barrington-Leigh, C. (2008). *Weather as a transient influence on survey-reported satisfaction with life. Draft research paper*. University of British Columbia.

Becker, W. E., & Watts, M. (1999). How departments of economics should evaluate teaching. *American Economic Review (Papers and Proceedings), 89*, 344–349.

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*, 709–719.

Brown, B. W., & Saks, D. H. (1987). The microeconomics of the allocation of teachers' time and student learning. *Economics of Education Review, 6*, 319–332.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*, 409–432.

Connolly, M. (2013). Some like it mild and not too wet: The influence of weather on subjective well-being. *Journal of Happiness Studies, 14*, 457–473.

De Giorgi, G., Pellizzari, M., & Redaelli, S. (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics, 2*(2), 241–275.

De Giorgi, G., Pellizzari, M., & Woolston, W. G. (2012). Class size and class heterogeneity. *Journal of the European Economic Association, 10*, 795–830.

De Philippis, M. (2013). *Research incentives and teaching performance evidence from a natural experiment. Mimeo*.

Denissen, J. J. A., Butalid, L., Penke, L., & van Aken, M. A. (2008). The effects of weather on daily mood: A multilevel approach. *Emotion, 8*, 662–667.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review, 102*, 1241–1278.

Figlio, D. N., & Kenny, L. (2007). Individual teacher incentives and student performance. *Journal of Public Economics, 91*, 901–914.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review (Papers and Proceedings), 100*, 250–255.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources, 14*, 351–388.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. In E. A. Hanushek, F. Welch (Eds.), *Handbook of the economics of education* (Vol. 1, pp. 1050–1078). Amsterdam: North Holland.

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

Hirsch, J. E. (6572). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 16569–16572.

Hogan, T. D. (1981). Faculty research activity and the quality of graduate training. *Journal of Human Resources, 16*, 400–415.

Holmstrom, B., & Milgrom, P. (1994). The firm as an incentive system. *American Economic Review, 84*, 972–991.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*, 101–136.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York, NY: Springer-Verlag.

Kane, T. J., & Staiger, D. O. (and Staiger, 2008). *Estimating teacher impacts on student achievement: An experimental evaluation. Technical Report 14607 NBER Working Paper Series*.

Keller, M. C., Fredrickson, B. L., Ybarra, O., Coté, S., Johnson, K., Mikels, J., et al. (2005). A warm heart and a clear head. The contingent effects of weather on mood and cognition. *Psychological Science, 16*, 724–731.

Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*, 59–63.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics, 114*, 497–532.

Lavy, V. (2009). Performance pay and teachers' effort, productivity and grading ethics. *American Economic Review, 95*, 1979–2011.

Mullis, I. V., Martin, M. O., Robitaille, D. F., & Foy, P. (2009). *TIMSS Advanced 2008 International Report*. Chestnut Hills, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

OECD. (2008). *Education at a glance*. Paris: OECD Publishing.

OECD. (2010). *PISA 2009 at a glance*. Paris: OECD Publishing.

Prendergast, C., & Topel, R. H. (1996). Favoritism in organizations. *Journal of Political Economy, 104*, 958–978.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review (Papers and Proceedings), 94*, 247–252.

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review (Papers and Proceedings), 100*, 261–266.

Rothstein, J. (2009). Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*, 537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*, 175–214.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513–523.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics, 87*, 355–374.

Swamy, P. A. V. B., & Arora, S. S. (1972). The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica, 40*, 261–275.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review (Papers and Proceedings), 100*, 256–260.

Weinberg, B. A., Fleisher, B. M., & Hashimoto, M. (2009). Evaluating teaching in higher education. *Journal of Economic Education, 40*, 227–261.

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business, 78*, 313–317.

Quality Assurance in Education

Students' perceptions of the evaluation of college teaching

Larry CrumbleyByron K. HenryStanley H. Kratchman

Article information:

To cite this document:
Larry CrumbleyByron K. HenryStanley H. Kratchman, (2001),"Students' perceptions of the evaluation of college teaching", Quality Assurance in Education, Vol. 9 Iss 4 pp. 197 - 207
Permanent link to this document:
http://dx.doi.org/10.1108/EUM0000000006158

Users who downloaded this article also downloaded:

For Authors

About Emerald www.emeraldinsight.com

Students' perceptions of the evaluation of college teaching

Larry Crumbley
Byron K. Henry and
Stanley H. Kratchman

The authors

Larry Crumbley is KPMG Endowed Professor at Louisiana State Unviersity, Baton Rouge, Louisiana, USA.
Byron K. Henry is an Assistant Professor, Howard University, Washington, DC, USA.
Stanley H. Kratchman is Professor of Accounting, Texas A&M University, College Station, Texas, USA.

Keywords

Evaluation, Higher education, Customer satisfaction

Abstract

The validity of student evaluation of teaching (SET) has been continually debated in the academic community. The primary purpose of this research is to survey student perceptions to provide any evidence of inherent weaknesses in the use of SETs to measure and report teaching effectiveness accurately. The study surveyed over 500 undergraduate and graduate students enrolled in various accounting courses over two years at a large public university. Students were asked to rate several factors on their importance in faculty evaluations and identify instructor traits and behaviors warranting lower ratings. The study provides further evidence that the use of student evaluations of teaching for personnel decisions is not appropriate. Students will punish instructors who engage in a number of well-known learning/teaching techniques, which encourages instructors to increase SET scores by sacrificing the learning process. Other measures and methods should be employed to ensure that teaching effectiveness is accurately measured and properly rewarded. Using student data as a surrogate for teaching performance is an illusionary performance measurement system.

Electronic access

The research register for this journal is available at
http://www.mcbup.com/research_registers

The current issue and full text archive of this journal is available at
http://www.emerald-library.com/ft

Quality Assurance in Education
Volume 9 · Number 4 · 2001 · pp. 197–207
© MCB University Press · ISSN 0968-4883

Introduction

Based on a review of policies and practices in evaluation, Seldin (1993) noted an 86 percent use of systematic student ratings to support curriculum and personnel/decisions in higher education, up from 68 percent in 1984, and 28 percent in 1973 (Seldin,1984). Moreover, Calderon *et al.* (1994) noted that 82 percent of accounting administrators use multiple information sources in assessing teaching performance, but 18 percent utilize only student evaluations. Thus, there has been a subtle power shift in the control of higher education from professors to their students. Reckers (1995, p. 33) states that "nearly 75 percent of academics judge student course evaluations as unreliable and imprecise metrics of performance, yet nearly 100 percent of schools use them, frequently exclusively."

The student evaluation of teaching (SET) questionnaire is a control device used to audit professors and/or measure their performance by students. This SET system causes professors to manipulate students and students in turn to manipulate teachers, and all of this symbiotic game playing is really about one thing: grades (Sacks, 1996). In order to stop the dwindling number of students in their classrooms, many instructors will inflate their grades and lighten their course assignments and tests (Bauer, 1996; Handlin, 1996). This use of SET data fuels the inappropriate attitudes that more and more students display, and makes grade inflation inevitable (Bauer, 1996).

In other areas where stakeholders rely on financial and other statistical information, they are deeply concerned with the relevance, reliability[1], comparability and neutrality of the data[2]. For example, the rise of the accounting profession occurred primarily because of the desire for better financial reporting (Carey, 1970). Both public and private institutions utilize external or internal audits, but few seem concerned with the relevance, reliability, comparability, and neutrality of the data used to evaluate teaching. Because administrative decisions are often irreversible, positions and even careers may be at stake, so information on which such decisions are made should be of proven validity. Not only has the validity of student ratings not been substantiated, but also more current empirical evidence has shown that student evaluations are misleading and/or invalid.

Many universities publish SET results in the library or on the Internet[3]. Once SET data is disclosed, stakeholders are confronted with the potential dangers of bias, misinterpretation, inexactness and ambiguity – the same dangers facing financial statements. Thus, there must be a body of theory that is generally accepted and universally practised. Currently most schools, departments, and colleges have a different theory and practice, so comparability is nearly impossible. Published SET data can lead to defamation of faculty reputation and may not meet what is considered contractual informed consent of faculty for their use (Haskell, 1997, p. 17).

Imagine CPA firms sending questionnaires to all of the clients of their employees asking the clients to evaluate the employees – and then publishing the results. Or imagine the IRS sending questionnaires to the taxpayers of each agent and then publishing the results. The inflow of revenue to the federal government would decrease over time. Imagine if the merit pay and firing practices of airline pilots were to be based on questionnaire data gathered from passengers at the end of each flight (with results published). More flights would be on time, but safety would be compromised.

The cause of grade inflation in higher education can be demonstrated by the following analogy. Suppose 200 people work at the end of an assembly line to reject defective products. Their pay increases, promotions, and psychological well-being are based on the total number of parts (both defective and non-defective) that reach the shipping docks. Assume that the total number of products manufactured and the actual number of defective parts are constant. Project over time what would happen to the total number of rejected products. This same situation may be occurring in higher education with the SET driven environment, and its own form of social promotion through grade inflation. In a faculty survey one-third of the respondents stated that they had substantially decreased the level of difficulty and grading standards for their courses (Ryan *et al.*, 1980).

Sources of invalidity

Numerous sources of invalidity in student evaluations of college teaching have been noted. Sheenan (1975) suggests several potential sources of invalidity:

- the extent to which student ratings reflect effective instruction;
- the construction of the rating instrument; and
- the susceptibility of student ratings to variations in instructions and to both subtle and overt instructor influence tactics.

Rodin and Rodin (1972, p. 1164) noted that there are two ways to use students in the evaluation of college teaching. The objective approach uses student opinions as a surrogate to measure teaching performance. Rodin and Rodin (1972) found a significant negative correlation between what students learned and their evaluation of teaching. These researchers concluded that students may resent instructors that force them to work hard and to learn more than students wish. Rodin and Rodin (1972, p. 1166) concluded that students were less than perfect judges of teaching effectiveness, and if student learning is a key component of "good teaching", the "good teaching" could not be validly measured by using student evaluations.

Variation in student ratings are attributed to class size (Meredith, 1984; Toby, 1993), gender of instructor and student (Basow and Silberg, l987; Sidanius and Crane, 1989) student classification and age, subject matter and course content, student abilities, achievements and expectations (Perkins *et al.*, 1990). Haladyna and Hess (1994) found that bias does exist and is a threat to valid interpretations and uses of SET data. Further, the use of simple summated ratings does not compensate for bias. "While increasing reliability diminishes random error, it does not eliminate constant error created by bias" (Haladyna and Hess, 1994). According to Ellis (1985), the SET instrument is biased, the whole process is biased, and the bias gets in the way of making effective use of student information.

Empirical evidence on the gender effect of faculty evaluations has been inconsistent. Feldman (1993) published a meta-analysis of 39 studies in the USA and Canada involving this gender effect. When significant differences were found in the studies, women were rated higher than men (p. 153). Further, students tend to rate same-gendered teachers a little higher than opposite-gendered teachers (p. 153). Kay (1979) found that students' anticipated grades in courses were associated with the gender effect and the overall evaluations of teachers. For male teachers, the higher the anticipated grade in the class, the higher both the male and female students' overall evaluation of the male teacher. However, the relationship was curvilinear for female students evaluating female teachers. Female students expecting a "B" evaluated the female teacher higher than female students expecting either an "A" or "C".

Langbein (1994) found that the effects of gender interacted with expected grades to impact student ratings. When the instructor exhibited behaviors inconsistent with traditional gender roles, the instructor received lower student evaluations. Because females were expected to be more nurturing and supportive, students that received lower than expected grades, penalized female instructors. Langbein (1994) concluded that the study supported the theoretical expectation that students treat female faculty members differently from otherwise comparable male faculty.

In an experiment Freeman (1994) found that both female and male students prefer teachers who possessed both feminine and masculine characteristics, regardless of the gender of the teacher. With a larger percentage of women (52.6 percent) versus men (47.4 percent) in many undergraduate accounting programs in 1994, this gender effect may be significant (Nelson and Deines, 1995).

Perkins *et al.* (1990) investigated the influence of grading standards and assigned grades on expected grades and student ratings of instruction. These authors (p. 641) concluded that whether grades were assigned randomly or in accord with student performance, there was evidence that students' evaluation of instruction were sensitive to grades professors assigned. Further, they observed that grade discrepancies (i.e. differences between assigned and expected grades) might have a strong effect on students' rating of teaching. Moreover, Johnson and Christian (1990) noted that expected grades were more highly correlated than assigned grades with student ratings. An explanation was that students did not know final grades at the time the evaluations were administered. In general, both studies found that students with higher than expected grades gave higher ratings than students with lower than expected grades.

Brown (1976, p. 577) observed that grades accounted for only 9 percent of variation in student ratings, but grades were substantially more influential than other factors expected to correlate with student ratings, such as class size, course level, gender and age of student, and instructor experience, measured by title and number of years teaching. More recently, however, Greenwald (1997) found that grades distort ratings (away from valid measures of instructional quality) by amounts corresponding to 9 percent to 20 percent of ratings variance. Two levels of contamination correspond to grades-ratings correlations of 0.30 and 0.45, respectively. Noting high correlations between grades and student ratings, Johnson and Christian (1990, p. 480) commented that if an instructor desired high evaluations, all he or she had to do was assign higher grades to students. Centra and Creech (1976) found in a sample of 14,023 students that students expecting an "A" grade gave a mean rating of 3.95, those expecting a "B", 3.74, those expecting a "C", 3.41, and those expecting a "D" grade gave a mean rating of 3.02. Expected grade also correlated with ratings of the value of the course to the student; individual student responses correlated 0.26 and the mean correlated 0.31. Basically, students give high ratings in appreciation for high grades (Aronson and Linder, 1965; Goldman, 1993). At least faculty believe that lenient grading produces higher student ratings and faculty reacts accordingly (Martin, 1998; Powell, 1977; Stumpf and Freedman, 1979; Winsor, 1977; Worthington and Wong, 1979; Yunker and Marlin, 1984).

While Frey *et al.* (1975, p. 440) did not find that grades varied systematically with ratings, the researchers observed that experienced students were clearly more lenient in their ratings than their younger colleagues. This finding suggests that instructors that teach upper-division or graduate level courses should receive higher ratings. These studies illustrate the impact that certain non-instructional factors may have on student ratings. Consequently, some instructors must overcome barriers to teaching on walking into the classroom. This study suggests that administrators must consider these barriers and bias if student evaluations are used in personnel decisions.

As previously noted, a possible source of invalidity is the susceptibility of ratings to instructor influence tactics (Sheenan, 1975). In an attempt to maximize ratings, Crumbley (1995) noted that instructors will inflate grades, reduce course content, and simplify examinations (i.e. impression management). Martin (1998) refers to "gaming", attempts by faculty to influence student ratings in their favor by using practices that distract from learning rather than enhance learning. Moreover, Renner (1981, p. 130) commented that student evaluations have diminished the quality of higher education by forcing instructors to target the wider range of average students instead of providing intellectual challenges for more able students. Brown (1976) remarked that student ratings may be counterproductive in education, by unfairly rewarding not the best teachers, but merely the most lenient. Hocutt (1987-1988) concludes that instructors can buy ratings with grades. When one professor lowers his "price", so must another professor, or lose out in the competition. The resulting paradox is professors competing with each other to give higher grades to students who are learning less: grade inflation and a corruption of the pedagogical purpose.

Student perceptions

Most studies have dealt with the quantitative aspects and have ignored such behavioral aspects as impression management, self-presentation, lying, and dysfunctional behavior. Also, the placement of such authority in the hands of students may be problematic for other reasons, and there is a paucity of research from the students' perception. Dwinell and Higbee (1993) surveyed over 150 students to gather opinions regarding the value of teaching evaluations. They observed that 67 percent of the respondents believed that rating forms were an effective means of evaluating instructors, and 83 percent believed that instructors changed their behavior as a result of weaknesses identified by the evaluations. However, the respondents did not believe that their evaluations affected faculty salaries, or promotion and tenure decisions. Dwinell and Higbee (1993) concluded that students may not understand the importance of evaluations, because the instructions do not specify how evaluation results will be used. Thus, an important question is what variables do students use to evaluate professors? This article looks at the SET system from the student's perspective.

Research design

Individual students enrolled in accounting courses over two years serve as the units of analysis for this study. Questionnaires were administered to various sections and courses offered by the accounting department at a large south-west university. The School of Business and the Accounting Department would be considered SET-driven (Crumbley, 1995). A total of 530 students were asked to rate the importance of 18 factors on instructor ratings on a five-point Likert-type scale, and the degree of agreement/disagreement with 17 statements related to instructor traits and behaviors that would lower overall teaching evaluation scores.

Differences in student perceptions based on student gender, student classification, and course grade expectations are analyzed using analysis of variance (ANOVA) and multiple comparison procedures. Bivariate correlation analyses were performed to draw further inferences regarding factors expected to impact student ratings.

Results

Respondent demographics are reported in Table I. Although 530 usable questionnaires were returned, some students did not respond to every question. The sample is primarily composed of upper-division students (i.e. seniors and graduate students). A total of 20 percent of the respondents indicated that they had cheated on at least one college examination. Would these same students, therefore, lie on SET questionnaire? Of the respondents, 36 percent indicated that they checked prior grade distributions for the purpose of choosing instructors.

A more interesting finding was that 36 percent of the respondents indicated that they would respond differently to teaching

Table I Respondent demographics

Gender		
Male	72	0.431
Female	95	0.569
Student classification		
Freshman	0	0.000
Sophomore	23	0.043
Junior	20	0.038
Senior	404	0.764
Graduate	82	0.155
Expected grade		
A	272	0.516
B	157	0.298
C	77	0.146
D	21	0.040
E	0	0.000
Current overall grade point average (GPA)		
3.50-4.00	144	0.273
3.00-3.45	163	0.309
2.50-2.99	183	0.347
Below	38	0.072
Have you ever cheated on an examination?		
Yes	107	0.202
No	423	0.798
Have you checked prior grade distributions of instructors?		
Yes	189	0.364
No	330	0.636
Would you respond to teaching evaluations differently if they were not anonymous?		
Yes	179	0.362
No	315	0.638

evaluations if they were not anonymous. While the anonymity of student responses may encourage honest assessments, it also allows students to "attack" instructors without fear of punishment. Instructors are not able to seek restitution from students responsible for damaging professional reputations. Furthermore, anonymity precludes the validation of student ratings. Only general student characteristics, such as sex and classification, can be investigated to ascertain influential components in the student evaluation process.

Percentages and mean responses for test items are reported in Tables II and III. Overall, students found teaching style (88.8 percent), presentation skills (89.4 percent), enthusiasm (82.2 percent), preparation and organization (87.3 percent), perceived learning (82.5 percent), and fair grading (89.8 percent) at least very important to rating instructors. Students also indicated that they were more apt to lower evaluations if instructors did not teach them enough to maintain their grade average (46.5 percent), asked embarrassing questions (41.9 percent), or appeared inexperienced (41.0 percent). An interesting observation was that students agreed that they would lower ratings if the instructor was white (13.3 percent) or male (13.1 percent). This result is interesting given the composition of the university and college student populations sampled (few minority students). If such bias exist, they must be corrected if student ratings of teaching effectiveness are to provide a basis for comparison. Some, but not all, of the male bias might be explained by the high composition of females in the classes (about 56 percent), and the fact that only courses taught by male instructors were included in the survey.

Respondents indicated that they could discern the impact of experience on classroom instruction. Of the respondents, 70 percent indicated that they would not lower ratings because their instructor was on a non-tenure track position, but 40 percent would lower ratings if the instructor appeared inexperienced.

Mean responses by gender and student classification are reported in Table IV. Although the results do not provide evidence of statistically significant gender or classification effects, the findings tend to support the general perception that opinions differ across factors beyond the control of college faculty[4]. Grading appears to be slightly more important to females than males, and to undergraduates relative to graduate students. The only significant difference noted between graduate and undergraduate students indicated that graduate students placed greater weight on class organization skills which are deemed important to successful task completion. Another possible explanation is that older students may prefer more structure than younger students.

Importance of grading and difficulty of instructor

From the point of view of these students, the grading of an instructor is extremely

Table II Percentages and means of students's responses

Question and content	Rank	1 Not at all important	2 Slightly important	3 Moderately important	4 Very important	5 Extremely important	Mean
4r Fair grader	1	0.4	1.9	7.9	37.4	52.4	4.39
4a Teaching style	2	0.2	1.3	9.7	46.4	42.4	4.29
4b Presentation skills	3	0.4	1.1	9.1	49.1	40.3	4.27
4m How well prepared and organized	4	0.4	1.5	10.8	48.2	39.1	4.24
4o How much I learned	5	0.6	1.9	15.1	42.2	40.3	4.19
4d Enthusiasm	6	0.4	1.7	15.7	47.8	34.4	4.14
4c Instructor's grading policy	7	1.5	5.5	26.3	38.6	28.2	3.86
4p Instructor's availability	8	0.8	8.5	25.7	39.5	25.5	3.80
4n How nice he/she is	9	1.1	11.3	31.9	34.2	21.4	3.63
4j Heavy class workload	10	5.5	10.0	28.0	38.4	18.1	3.53
4i Tough grading	11	5.3	14.0	39.5	28.5	12.7	3.29
4e Course difficulty	12	4.5	14.2	43.7	25.1	12.5	3.26
4l Raise the level of course	13	21.4	23.3	28.6	21.2	5.5	2.66
4g Class size	14	20.4	28.0	30.1	13.4	8.1	2.60
4k Major vs non-major course	15	26.5	21.9	29.7	14.9	7.0	2.54
4f Time of day when course is taught	16	35.8	20.1	24.1	12.7	7.4	2.35
4b Whether required or elective course	17	34.2	24.6	25.5	10.8	4.9	2.27
4q Giving students free time	18	43.4	29.2	18.7	4.5	4.2	1.96

Table III Percentages and means of students' responses, percentages and means of factors likely to lower student evaluation of teaching (SET) scores

Question and content	Rank	1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree	Mean
Percentages and means of students' responses							
5a (Instructor) has not taught me enough to make a grade of at least my current GPA	1	9.7	19.3	24.4	11.7	34.8	3.19
5b Asks me embarrassing questions	2	16.8	19.8	21.4	13.0	28.9	3.01
5q Appears to be inexperienced	3	17.3	17.3	24.4	9.3	31.7	2.98
5d Is a decent teacher, but grades very hard	4	9.5	34.2	28.2	4.2	24.0	2.78
5l Requires a significant amount of homework	5	15.4	30.1	33.5	3.1	17.9	2.63
5m Gives pop quizzes (i.e. unannounced)	6	21.4	28.7	22.7	4.6	22.5	2.60
5o Introduces religion into the course	7	32.4	13.9	29.9	10.0	13.7	2.55
5p Introduces politics into the course	8	25.9	19.9	34.6	6.0	13.7	2.54
5c Uses overhead materials, are not typed	9	29.7	36.7	19.3	4.7	9.6	2.23
Percentages and means of factors likely to lower student evaluation of teaching (SET) scores							
5j Calls on students randomly in class	10	32.2	32.4	24.3	3.1	8.1	2.17
5n Takes up and grades homework	11	32.6	34.3	24.1	1.7	7.3	2.11
5e He/she is an instructor or graduate assistant (i.e. not a professor)	12	38.8	31.9	21.7	2.1	5.5	2.00
5k Uses humor	13	62.0	15.4	8.1	8.1	6.4	1.83
5h Is white (i.e. ethnic majority)	14	68.3	8.3	10.1	13.1	0.2	1.81
5g Is male	15	68.0	9.6	9.2	12.9	0.2	1.80
5i Is non-white (i.e. minority or foreign)	16	79.6	7.7	10.9	1.0	0.8	1.35
5f Is female	17	80.7	10.0	8.9	0.4	0.0	1.29

important. At least 36 percent of the students had checked the prior grade distributions of instructors. At this university all grade distributions were made available to students in an office open the entire week, and the information could be obtained online by computer.

Of the 18 factors in Table II, "fair grader" was ranked first (mean of 4.39) with "how much I learned" a distant fifth (mean of

Table IV Means of students' responses by group

		By gender		By class	
Question and content	Rank	Male	Female	Graduate	Undergraduate
4r Fair grader	1	4.38	4.48	4.25	4.42
4a Teaching style	2	4.22	4.33	4.45	4.26
4b Presentation skills	3	4.30	4.27	4.37	4.26
4m How well prepared and organized	4	4.19	4.36	4.47	4.19
4o How much I learned	5	4.18	4.10	4.21	4.12
4d Enthusiasm	6	4.15	4.11	4.39	4.16
4c Instructor's grading policy	7	3.98	3.88	3.75	3.88
4p Instructor's availability	8	3.84	3.87	3.85	3.79
4n How nice he/she is	9	3.59	3.53	3.62	3.63
4j Heavy class workload	10	3.59	3.49	3.21	3.59
4i Tough grading	11	3.29	3.34	2.96	3.35
4e Course difficulty	12	3.33	3.22	3.10	3.29
4l Higher the level of course	13	2.65	2.38	2.54	2.68
4g Class size	14	2.54	2.48	2.56	2.61
4k Major vs non-major course	15	2.59	2.34	2.12	2.61
4f Time of day when course is taught	16	2.51	2.22	2.07	2.41
4h Whether required or elective course	17	2.27	2.16	1.96	2.33
4q Giving students free time	18	1.90	1.86	1.84	1.99

4.19), followed by "instructor's grading policy"at seventh (mean of 3.86). Further "heavy class load" was ranked tenth (mean of 3.53), and "tough grading" was ranked 11th (mean of 3.29).

With respect to factors likely to lower a student's evaluation of an instructor, five of the first six items dealt with grades or tough professors:

(1) Not taught enough to make expected grade, ranked 1.
(2) Ask me embarrassing questions, ranked 2.
(3) Grades very hard, ranked 4.
(4) Significant homework, ranked 5.
(5) Gives pop quizzes, ranked 6.

From a different perspective, many of the techniques used to help educate students may be used by students to punish instructors (see Table V).

In other words, 42 percent of the students will punish instructors for being asked embarrassing questions, 28 percent for being graded hard, 27 percent for pop quizzes, 20 percent for significant homework, 14.5 percent for using humor, 14.3 percent for use of untyped overheads, 11 percent for being called on and 9 percent for merely grading homework. At the same time, many students will reward a professor for being perceived as nice (55.6 percent in Table II). The safest approach for an instructor is to lecture, be nice, grade easy, and cover little material. Apparently, learning in a SET-driving university may not be important to a significant number of students (learning was ranked only fifth in Table II). An instructor may use Table III to improve SET scores – especially by sacrificing the learning process. Also, an instructor should avoid religious remarks (23.7 percent punishment factor) and political comments (19.7 percent punishment factor) during the semester (i.e. be politically correct).

Table V Teaching techniques for which students may punish instructors

	Percentage impacted (%)	Ranked by mean
Asking embarrassing questions	4.9	2
Grades hard	28.2	4
Pop quizzes	27.1	6
Significant homework	20.0	5
Uses humor	14.5	13
Untyped overheads	14.3	9
Calling on students	11.2	10
Grading homework	9.0	1

Invalidity due to lying

Determining whether students punish instructors for teaching attributes that aid in learning depends on self-reporting. In order to test for self-reporting bias, question 9 asked: "If you call an employee stupid, the person will not evaluate you highly on an anonymous questionnaire". Although somewhat ambiguous, the authors expected an overwhelming "yes" response. An overwhelming 86.1 percent of the students answered "yes" to this question, with only 13.9 percent selecting "no." When this same question was introduced in subsequent classes *alone*, the students gave 93 percent "yes" response and only a 7 percent "no" response. In both cases, students recognized the harmful effect of a hard grading instructor.

Pilcher (1994, p. 88) states that the use of grades has caused most students to reluctantly comply to doing whatever it takes to receiving the grades that will reward them or keep them from receiving negative consequences. Hocutt (1887-1888) believes students desire SETs so they can control the behavior of professors. So do students lie on SET questionnaires in order to punish tough instructors? DePaulo *et al.* (1996) reported that students lie in approximately one out of every three of their social interactions. Both students and the community as a whole told relatively more self-centered lies to men, and relatively more other-oriented lies to women (p. 1). Other-oriented lies were told to protect or enhance other persons psychologically, or to advantage or protect the interest of others.

DePaulo *et al.* (1996) break lies into two major categories:

(1) self-centered; and
(2) other-oriented.

Self-centered lies are told to protect or enhance the liars psychologically, or to advantage or protect the liars' interests. This 7 percent difference could be self-centered lies, which is confirmed by the results in question 10. Kashy and DePaulo (1996) found that students that are more manipulative tell more lies, and that students who are more highly concerned with impression management tell more lies.

Only 20 percent of the students admitted to cheating on examinations even once. This percentage is quite low when compared with much higher percentages in other surveys. In one study 56 percent of a group of upper-level accounting students in four large public universities admitted to cheating on a test, project, or written assignment in college (Ameen *et al.*, 1995). If a student will cheat on an examination, can the student be expected to give an honest evaluation of an instructor?

Thus, if these students lied on questions 9 and 10, did they lie on the other questions involving grading and difficulty of instructors for self-centered purposes? If so, the punishment factor could be much higher. How can SET data be valid when there is no control and measurement of response bias for most questionnaires?

This dishonesty may explain why higher level courses receive higher SET scores (Arreola, 1994). In upper level courses, grades are higher, classes are smaller, and the instructor moves closer to friend or mentor (away from a stranger). Similarly, SET scores in large classes tend to be lower (Cashin and Slawson, 1977). In large classes an instructor has more difficulty becoming a "friend" of each student.

Summary and recommendations

Researchers should expect a range of responses on SET questionnaires, since most instructors will not be equally effective with all students (Dwinell and Higbee, 1993). However, ratings are illusionary when response differences cannot be attributed solely to instruction. Using SET data in performance evaluation is not an effective proxy for measuring student learning. When 42 percent of the students will punish a professor for embarrassing them because they have not prepared their homework or prepared for class, tough and difficult instructors are placed at a severe disadvantage in a SET-driven environment. Grading hard, giving pop quizzes and significant homework, and using humor should not be penalized in a performance evaluation. Dysfunctional behaviors spawned by the use of SETs for personnel decisions must be addressed in order to maintain credibility in the grading system, to better align performance with rewards, and to safeguard the integrity of the learning process.

Student evaluation of teaching serves many purposes. Evaluations are used for making personnel decisions, allocating faculty resources, diagnosing and improving teaching performance, and choosing courses and instructors. The objectives of the teaching evaluation process must be clearly defined. If student evaluations are used only to provide feedback to faculty members, then they need not be validated. However, if rewards and penalties are to be assigned, it is important that the rating measure what it purports to measure. SET data is a poor surrogate for measuring teaching performance. If the SET data is disclosed to the public, a university or college could be exposed to defamation lawsuits.

If the use of student evaluations in personnel decisions is continued, administrators should develop better instruments to measure teaching performance or correct raw student ratings to remove the effects of non-instructional bias[5]. Credibility of SET data rests on its reliability and neutrality. To be useful SET data must be trustworthy. Although higher education is a multi-billion dollar business, few attempts have been made to provide evenhanded, neutral, and unbiased data for evaluating teaching.

Player handicaps are provided in such sports as golf and bowling. In the case of faculty evaluations, handicaps may be appropriate in order to level the playing field for those inherently biased by the student evaluation process. Renner (1981) suggested that students rate the course, not the instructor, in order to place students in a less authoritative role. Given numerous potential sources of invalidity and bias, the continued use of student evaluations in personnel decisions is questionable indeed. One administrator (Hocutt, 1987-1988) states that SETs were instituted because students wanted them and because universities wanted students. Students did not want them because "they hanker for teachers from whom they would learn; they wanted them because they hungered for teachers whose behavior they could not control. To keep enrolment up, universities gave in to their students, while declaring their desire to improve teaching."

Unless a university or college has all regulated classes (common examinations, common curves, etc.), a SET-driven climate will result in severe grade inflation over time because of the natural behavior of the two actors – professor and student. Many of the students will select those courses where there is a greater possibility of receiving a higher grade. Other students will punish professors for grading hard. As a result, a significant number of professors engage in dysfunctional techniques (i.e. anti-learning) which causes continuous upward spiral in the average grades (e.g. a ratchet effect).

Notes

[1] For example, in December 1995, the SET questionnaires at the university from across the campus were stored unsecured on a table in a foyer on the second floor of a busy building. Students perform much of the work on the tabulation. How can SET data be valid and reliable when there is no security?

[2] During just one week in late January 1996, the online SET reports at this university were viewed 808 times. During three-and-a-half months in 1995, 3,342 reviewed the GPA data of the professors, but there was no record of how many students viewed the grade point data online.

[3] Authorities are beginning to question the legality of publicly releasing SET data (Robinson *et al.*, 1996).

[4] The researchers noted that only 167 (out of 530 respondents) were asked to identify their gender. Moreover, the sample is heavily skewed by the proportion of undergraduate students (85 percent). Both conditions may reduce the power of the procedures applied to detect differences between groups.

[5] Greenwald (1997) states that grading policies (combined relative and absolute grade) causes 27.3 percent contamination and class level (freshman to graduate) and enrollment cause 3.6 percent contamination. Thus, on a scale of 5.0, a professor could be penalized as much as 1.55.

References

Ameen, E.C., Guffey, D.M. and McMillan, J.J. (1995), "Accounting students' perceptions of questionable academic practices and factors affecting their propensity to cheat", paper at Southeast Regional AAA, pp. I-2B.

Aronson, G. and Linder, D.E. (1965), "Gain and loss of esteem as determinants of interpersonal attractiveness", *Journal of Experimental Social Psychology*, Vol. 1, pp. 156-71.

Arreola, R.A. (1994), *Developing a Comprehensive Faculty Evaluation System*, CEDA, Memphis, TN, p. 288.

Basow, S.A. and Silberg, N.T. (1987), "Student evaluations of college professors: are female and male professors rated differently?", *Journal of Educational Psychology*, Vol. 79 No. 3, 14 September, pp. 308-14.

Bauer, H.H. (1996), "The new generations: students who don't study," *The Technological Society at*

Risk Symposium, Orlando, FL, 10 September, pp. 1-37.

Brown, D.L. (1976), "Faculty ratings and student grades: a university-wide multiple regression analysis", *Journal of Educational Psychology*, Vol. 68 No. 5, pp. 573-8.

Calderon, T.G., Green, B.P. and Reider, B.P. (1994), "Extent of use of multiple information sources in assessing accounting faculty teaching performance", working paper.

Carey, J.L. (1970), *The Rise of the Accounting Profession 1937-1969*, AICPA, New York, NY, p. 1.

Cashin, W.E. and Slawson, H.M. (1977), *IDEA Technical Report No. 2: Description of Data Base*, Kansas State University, Center for Faculty Evaluation and Development, New York, NY.

Centra, J.A. and Creech, F.R. (1976), "The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness", *SIR Report No. 4*, Educational Testing Service, Princeton, NJ, pp. 24-7.

Crumbley, D.L. (1995), "The dysfunctional atmosphere of higher education: games professors play", *Accounting Perspectives*, Spring, pp. 67-76.

DePaulo, B.M., Kashy, D.A. and Ansfield, M.E. (1996), "Lying in relationship", paper presented at the 103rd meeting of the American Psychological Association, August, New York, NY.

Dwinell, P.L. and Higbee, J.L. (1993), "Students' perceptions of the value of teaching evaluations", *Perceptual and Motor Skills*, Vol. 76, pp. 995-1000.

Ellis, R. (1985), "Ratings of teachers by their students should be used wisely – or not at all", *The Chronicle of Higher Education*, Vol. 20 No. 31, November, p. 88.

Feldman, K.A. (1993), "College students' view of male and female college teachers: part II – evidence from students' evaluations of their classroom teachers", *Research in Higher Education*, Vol. 34 No. 2, pp. 151-91.

Freeman, H.R. (1994), "Student evaluation of college instructors: effects of type of course taught, instructor gender and gender role, and student gender", *Journal of Educational Psychology*, Vol. 86 No. 4, pp. 627-30.

Frey, P.W., Leonard, D.W. and Beatty, W.M. (1975), "Student ratings of instruction: validation research", *American Educational Research Journal*, Vol. 12 No. 4, pp. 435-47.

Goldman, L. (1993), "On the erosion of education and the eroding foundations of teacher education", *Teacher Education Quarterly*, Vol. 20, pp. 57-64.

Greenwald, A.G. (1997), "Validity concerns and usefulness of student ratings of instruction", *American Psychologist*, Vol. 52 No. 11, November, pp. 1182-7.

Haladyna, T. and Hess, R.K. (1994), "The detection and correction of bias in student ratings of instruction", *Research in Higher Education*, Vol. 35 No. 6, December, pp. 669-87.

Handlin, O. (1996), "A career at Harvard", *American Scholar*, Vol. 65 No. 5, Winter, pp. 47-58.

Haskell, R.E. (1997), "Academic freedom, tenure, and student evaluation of faculty: galloping polls in the twenty-first century", *Education Policy Analysis Archives*, Vol. 5 No. 6, pp. 1-32.

Hocutt, M.O. (1987-1988), "De-grading student evaluations: what's wrong with student polls of teaching", *Academic Questions*, Winter, pp. 55-64.

Johnson, R.L. and Christian, V.K. (1990), "Relation of perceived learning and expected grade to rated effectiveness of teaching", *Perceptual and Motor Skills*, Vol. 70, pp. 479-82.

Kay, S. (1979), "Sex bias in students' responses", *News for Teachers of Political Science: A Publication of the American Political Science Associatio*n, Vol. 23, pp. 17-19.

Kashy, D.A. and DePaulo, B.M. (1996), "Who lies", *Journal of Personality and Social Psychology*, Vol. 70 No. 5, pp. 1037-51.

Langebein, L.I. (1994), "The validity of student evaluations of teaching", *Political Science & Politics*, September, pp. 545-53.

Martin, J.R. (1998), "Evaluating faculty based on student opinions: problems, implications and recommendations from Deming's theory of management perspective", *Issues in Accounting Education*, Vol. 13 No. 4, pp. 1079-94.

Meredith, G.M. (1984), "Diagnostic and summative appraisal ratings of instruction", *Psychological Reports*, Vol. 46, pp. 21-2.

Nelson, I.T. and Deines, D.S. (1995), "Accounting student characteristics: results of the 1993 and 1994 Federation of Schools of Accountancy (FSA) surveys", *Journal of Accounting Education*, Vol. 13 No. 4, pp. 393-411.

Perkins, D., Gueri, D. and Schleh, J. (1990), "Effects of grading standards information, assigned grade, and grade discrepancies on student evaluations", *Psychological Reports*, Vol. 66, pp. 635-42.

Pilcher, J.K. (1994), "The value-driven meaning of grades", *Educational Assessment*, Vol. 2, February, pp. 69-88.

Powell, R.W. (1977), "Grades, learning, and student evaluation of instruction", *Research in Higher Education*, Vol. 7, pp. 193-205.

Reckers, P.M.J. (1995), "Know thy customer. Change in accounting education: a research blueprint", *Federations of Schools of Accounting*, pp. 29-35.

Renner, R.R. (1981), "Comparing professors: how student ratings contribute to the decline in quality of higher education", *Phi Delta Kappan*, October, pp. 128-30.

Robinson, R.K., Canty, A. and Fink, R.L. (1996), "Public disclosure of teaching evaluations: privacy and liability considerations", *Journal of Education for Business*, Vol. 71 No. 5, pp. 284-7.

Rodin, M. and Rodin, B. (1972), "Student evaluations of teachers: students rate most highly instructors from whom they learn least", *Science 177*, September, pp. 1164-6.

Ryan, J.J., Anderson, J.A. and Birchler, A.B. (1980), "Evaluations: the faculty responds", *Research in Higher Education*, Vol. 12 No. 4, pp. 317-33.

Sacks, P. (1996), *Generation X Goes To College*, Open Court, Chicago, IL.

Seldin, P. (1984), *Changing Practices in Faculty Evaluation*, Jossey-Bass, San Francisco, CA.

Seldin, P. (1993), "The use and abuse of student ratings of professors", *The Chronicle of Higher Education*, 12 June, p. A40.

Sheenan, D.S. (1975), "On the invalidity of student ratings for administrative personnel decisions", *Journal of Higher Education*, Vol. 46 No. 6, pp. 687-700.

Sidanius, J. and Crane, M. (1989), "Job evaluation and gender: the case of university faculty", *Journal of Applied Social Psychology*, Vol. 19 No. 2, pp. 174-97.

Stumpf, S.A. and Freedman, R.D. (1979), "Expected grade covariation with student ratings of instructors", *Journal of Educational Psychology*, Vol. 71, pp. 273-302.

Toby, S. (1993), "Class size and teaching evaluation", *Journal of Chemical Education*, Vol. 70 No. 6, pp. 465-6.

Winsor, J.L. (1977), "A's, B's, but not C's: a comment", *Contemporary Education*, Vol. 48, pp. 82-4.

Worthington, A.G. and Wong, P.T.P. (1979), "Effects of earned and assigned grades on student evaluations of an instructor", *Journal of Educational Psychology*, Vol. 71, pp. 764-75.

Yunker, J.A. and Marlin, J.W. (1984)," Performance evaluation of college and university faculty: an economic perspective", *Educational Administration Quarterly*, Winter, pp. 9-37.

This article has been cited by:

1. Chenicheri Sid Nair, Jinrui Li, Li Kun Cai. 2015. Academics' feedback on the quality of appraisal evidence. *Quality Assurance in Education* **23**:3, 279-294. [Abstract] [Full Text] [PDF]
2. Tang Howe Eng, Alif Faisal Ibrahim, Noor Emma Shamsuddin. 2015. Students' Perception: Student Feedback Online (SuFO) in Higher Education. *Procedia - Social and Behavioral Sciences* **167**, 109-116. [CrossRef]
3. Sami El-Mahgary, Petri Rönnholm, Hannu Hyyppä, Henrik Haggrén, Jenni Koponen, Steve Cook. 2014. Evaluating the performance of university course units using data envelopment analysis. *Cogent Economics & Finance* **2**, 918856. [CrossRef]
4. Zenawi Zerihun, Jos Beishuizen, Willem Os. 2012. Student learning experience as indicator of teaching quality. *Educational Assessment, Evaluation and Accountability* **24**, 99-111. [CrossRef]
5. Hilary Catherine Murphy, Harry de Jongh. 2011. Student perceptions of information system subject learning in hospitality management degree programmes. *International Journal of Contemporary Hospitality Management* **23**:3, 393-409. [Abstract] [Full Text] [PDF]
6. Donald Larry Crumbley, Ronald E. Flinn, Kenneth J. Reichelt. 2010. What is Ethical About Grade Inflation and Coursework Deflation?. *Journal of Academic Ethics* **8**, 187-197. [CrossRef]
7. Jenny A. Darby. 2008. Course evaluations: a tendency to respond "favourably" on scales?. *Quality Assurance in Education* **16**:1, 7-18. [Abstract] [Full Text] [PDF]
8. Ahmad Al-Issa, Hana Sulieman. 2007. Student evaluations of teaching: perceptions and biasing factors. *Quality Assurance in Education* **15**:3, 302-317. [Abstract] [Full Text] [PDF]
9. Alfred P. Rovai, Jonathan W. Kohns, Henry F. Kelly, Nancy E. Rhea. 2007. The Inquisition: A Parody for Christian Student Evaluation. *Christian Higher Education* **6**, 163-183. [CrossRef]
10. James S. Pounder. 2007. Is student evaluation of teaching worthwhile?. *Quality Assurance in Education* **15**:2, 178-191. [Abstract] [Full Text] [PDF]
11. Rod Gapp, Ron Fisher. 2006. Achieving excellence through innovative approaches to student involvement in course evaluation within the tertiary education sector. *Quality Assurance in Education* **14**:2, 156-166. [Abstract] [Full Text] [PDF]
12. Anna Maria TammaroQuality Assurance in Library and Information Science (LIS) Schools: Major Trends and Issues 389-423. [Abstract] [Full Text] [PDF] [PDF]
13. Hulya Julie Yazici. 2005. A study of collaborative learning style and team learning performance. *Education + Training* **47**:3, 216-229. [Abstract] [Full Text] [PDF]
14. Rosario Andreu, Lourdes Canós, Susana de Juana, Encarnación Manresa, Laura Rienda, Juan José Tarí. 2003. Critical friends: a tool for quality improvement in universities. *Quality Assurance in Education* **11**:1, 31-36. [Abstract] [Full Text] [PDF]

Routledge
Taylor & Francis Group

The Journal of Economic Education

ISSN: 0022-0485 (Print) 2152-4068 (Online) Journal homepage: http://www.tandfonline.com/loi/vece20

Determinants of How Students Evaluate Teachers

Michael A. McPherson

To cite this article: Michael A. McPherson (2006) Determinants of How Students Evaluate Teachers, The Journal of Economic Education, 37:1, 3-20, DOI: 10.3200/JECE.37.1.3-20

To link to this article: http://dx.doi.org/10.3200/JECE.37.1.3-20

Published online: 07 Aug 2010.

Research in Economic Education

In this section, the *Journal of Economic Education* publishes original theoretical and empirical studies of economic education dealing with the analysis and evaluation of teaching methods, learning, attitudes and interests, materials, or processes.

PETER KENNEDY, Section Editor

Determinants of How Students Evaluate Teachers

Michael A. McPherson

Abstract: Convincingly establishing the determinants of student evaluation of teaching (SET) scores has been elusive, largely because of inadequate statistical methods and a paucity of data. The author uses a much larger time span than in any previous research—607 economics classes over 17 semesters. This permits a proper treatment of unobserved heterogeneity. Results indicate that instructors can buy higher SET scores by awarding higher grades. In principles classes, the level of experience of the instructor and the class size are found to be significant determinants of SET scores. In upper-division classes, the type of student and the response rate matter. In both types of classes, factors specific to courses, instructors, and time periods are important; adjustments of scores to remove these influences may be warranted.

Key words: class size, student evaluations of teaching, unobserved heterogeneity
JEL code: A22

An extensive literature surrounds the issue of student evaluation of teaching (SET) scores. Research in this area began as early as 1936 with Heilman and Armentrout's article in the *Journal of Educational Psychology* and has continued unabated. The quantity of research is indicative of the importance of SET in higher education. For better or for worse, it is now standard for universities to

Michael A. McPherson *is an associate professor of economics at the University of North Texas (e-mail: mcpherson@unt.edu).* *The author is grateful for the assistance of David Molina and Caesar Righton in assembling the data and the insightful suggestions of Jeffrey Rous and R. Todd Jewell. The suggestions of three anonymous referees and Peter Kennedy were especially valuable.*

expect departments of economics to evaluate faculty, at least in part according to their SET scores (Becker and Watts 1999). The findings of researchers in this area are varied and sometimes in opposition to each other.[1] Unfortunately, statistical shortcomings and problems related to the data themselves have hampered much of the work in this area. My results add to the literature by addressing a critical statistical problem that has plagued previous work: unobserved heterogeneity. Although a few efforts have been made to tackle this problem (e.g., Mason, Steagall, and Fabritius 1995; Tronetti 2001), each has had other statistical shortcomings, such as a failure to test for endogeneity or a lack of time-series data.

In this study, I tested for endogeneity and controlled for unobserved heterogeneity and used a much longer time period than any previous work—17 semesters from 1994 to 2002. In the model, I employed controls for instructor, course, and semester-specific effects. The results suggested some obvious ways that rankings of instructors by SET scores might be adjusted. Clearly this is an important issue given the importance that such rankings have in determining such things as promotion, tenure, and merit raises. I examined whether adjustments of this nature would significantly alter the rankings of instructors.

MODEL AND DATA DESCRIPTION

I obtained the data from the University of North Texas's (UNT's) Academic Records Office and from the Department of Economics.[2] The data covered the 17 semesters between August 1994 and December 2002 and comprised 607 individual undergraduate classes taught by a total of 35 different instructors. It is possible that the relationships between SET scores and the explanatory variables differ for introductory economics classes compared with upper-division classes. Earlier researchers with access to data from both principles and upper-division courses routinely pooled the classes together without conducting tests of the appropriateness of such a grouping (see for example, Danielson and White 1976; Aigner and Thum 1986; Isely and Singh 2005). I conducted an *F* test for the appropriateness of pooling together principles courses with upper-division courses; these tests indicated that it is inappropriate to pool the data,[3] and, as a result I present regression results separately for principles and upper-division observations.[4] The principles subsample comprised 360 classes taught by 28 individual instructors. There were 247 upper-division classes, with 20 individual instructors. Both samples excluded classes with fewer than 15 students and included only instructors teaching at least three classes over the 17-semester time frame.

The instructors distributed SET forms without announcement beforehand[5] near the end of the semester and the forms were anonymous. The form included 25 questions, some of which were phrased in a positive and some in a negative manner. The answers were on 4-point scale with a 1 indicating *strong agreement* and a 4 indicating *strong disagreement* with the question. Given this instrument, there are many possible ways to measure quality of teaching. In this research, I used as the dependent variable (hereafter referred to as EVAL) an average of four questions: "I would take another course that was taught in this way"; "The instructor did not synthesize, integrate or summarize effectively"; "Some things were not explained very well"; and "I think that the course was taught quite well."[6] A second dependent variable consisted of the average value from the last of these four questions. Because the results differed only slightly, the results from regressions involving the second dependent variable are not presented here, although any substantial differences will be noted. In principle, EVAL can range from 1 to 4, with a 1 representing the best possible SET score and a 4 representing, at least from the students' points of view, poor teaching. For the principles classes sample, the average score for the dependent variable was (1.86) in Table 1. It is not surprising that evaluation scores were better for instructors of upper-division classes (1.74).

Following the literature, the determinants of the SET score are likely to fall into several categories. First are characteristics of the students in each class; these include such measures as major (PCTMAJ), expected grade (EXPGRADE), and the proportion of students who completed the evaluation questionnaire (RESPONSE).[7] PCTMAJ measures the percentage of the class that is majoring in economics; the average was 39.2 percent for upper-division classes and under 1 percent for principles classes. Economics majors might be more favorably disposed toward economics classes and instructors and perhaps better able than nonmajors to evaluate their economics instructors' abilities to teach economics, so SET scores might be better in such classes. I collected data on expected grades (EXPGRADE) as part of the evaluation exercise, and measured the variable on the usual 4-point scale, averaging 2.88 (principles classes) and 3.22 (upper-level courses) for these data. This effect has been of particular interest in the literature. Isely and Singh (2005) argued that it is the difference between expected grade and the grades that students are accustomed to receiving that matters. That is, when the average grade expected by a given class is above the average grade point average (GPA) of the class before the semester began, higher SET scores may result. However appealing this variable may be intuitively, it is not clear that such a variable can be properly calculated. The average expected grade was calculated from the evaluation exercise and, as such, represented only those students who were present on the day of the exercise and who chose to participate. The average GPA of the students responding to the evaluation questionnaire was not known, however. Instead, it was the average GPA of all students registered for the course that was available. Given that the distribution of students who did not participate in the evaluation process was unlikely to be the same as that of students who did, the validity of such a variable was questionable at best. In any event, I calculated this surprise variable in a similar manner to Isely and Singh—as the difference between expected grade and overall GPA. As I note later, a series of Davidson-MacKinnon *J* tests (Davidson and MacKinnon 1981) indicated that, in general, it was more appropriate to use EXPGRADE than this surprise variable. I included the response rate, measured as the percentage of the students registered who actually participated in the evaluation process, as a way to control for possible selection bias. In addition, this variable may be indicative of student interest, in which case it would be reasonable to expect it to have a beneficial effect on SET scores. Alternatively, a high response rate might mean that a larger number of poorly performing students were evaluating their instructors. This might have detrimental effect on an instructor's SET score. In any case, RESPONSE averaged about 68

TABLE 1. Descriptive Statistics

	Principles classes			Upper-division classes		
Variable	Mean (st. dev.)	Min.	Max.	Mean (st. dev.)	Min.	Max.
EVAL	1.86 (0.40)	1.25	3.77	1.74 (0.42)	1.05	3.13
EXPGRADE	2.88 (0.24)	2.31	3.80	3.22 (0.29)	2.33	3.92
PCTMAJ	0.58 (1.12)	0.00	7.69	39.18 (22.57)	0.00	100.00
ONEDAY	0.08 (0.26)	0.00	1.00	0.38 (0.49)	0.00	1.00
TWODAY	0.42 (0.49)	0.00	1.00	0.39 (0.49)	0.00	1.00
THREEDAY	0.51 (0.50)	0.00	1.00	0.23 (0.42)	0.00	1.00
CLSIZE	82.34 (44.33)	19.00	318.00	32.96 (8.89)	16.00	53.00
EXPERIENCE: 1 TO 4 SEMESTERS	0.37 (0.48)	0.00	1.00	0.27 (0.44)	0.00	1.00
EXPERIENCE: 5–10 SEMESTERS	0.35 (0.48)	0.00	1.00	0.37 (0.48)	0.00	1.00
EXPERIENCE: 11+ SEMESTERS	0.27 (0.45)	0.00	1.00	0.36 (0.48)	0.00	1.00
RESPONSE (rate)	67.89 (12.51)	26.74	97.87	69.89 (13.76)	33.33	100.00
ECON1100: Micro	0.38 (0.49)	0.00	1.00			
ECON1110: Macro	0.62 (0.49)	0.00	1.00			
ECON3000: Contemp. issues				0.02 (0.13)	0.00	1.00
ECON3050: Consumer				0.04 (0.21)	0.00	1.00
ECON3150: Discrimination				0.06 (0.24)	0.00	1.00
ECON3550: Micro				0.15 (0.36)	0.00	1.00
ECON3560: Macro				0.12 (0.33)	0.00	1.00
ECON4020: Money and banking				0.15 (0.36)	0.00	1.00
ECON4100: Comp. systems				0.03 (0.18)	0.00	1.00
ECON4140: Managerial				0.03 (0.18)	0.00	1.00
ECON4150: Public finance				0.06 (0.25)	0.00	1.00
ECON4180: Health				0.03 (0.18)	0.00	1.00
ECON4290: Labor				0.03 (0.17)	0.00	1.00
ECON4440: Environmental				0.04 (0.19)	0.00	1.00
ECON4460: Ind. organization				0.01 (0.09)	0.00	1.00
ECON4500: Sports				0.01 (0.11)	0.00	1.00
ECON4510: History of thought				0.03 (0.18)	0.00	1.00
ECON4600: Development				0.03 (0.17)	0.00	1.00
ECON4630: Research methods				0.01 (0.09)	0.00	1.00
ECON4850: Trade				0.07 (0.25)	0.00	1.00
ECON4870: Econometrics				0.07 (0.25)	0.00	1.00
Sample size	360			247		

percent and ranged from about 27 percent to 100 percent. The response rate for principles classes was not significantly lower than that of the upper-division sample.

A second group of possible determinants of SET are characteristics of the course, such as the level of the course, the length of the class period, the number of students in the class, and so forth. In particular, I modeled the level of the course using a series of dummy variables. In the case of the principles data, the base category was principles of macroeconomics (ECON1110). As shown in

Table 1, 62 percent of the principles classes in the data were principles of macroeconomics, with the remaining 38 percent, principles of microeconomics. With respect to the upper-division courses, there were 20 different courses students could take. In the regressions involving upper-division classes, intermediate micro (ECON3550) served as the base category. The most common upper-division classes were intermediate macro (ECON3560), intermediate micro (ECON3550), and money and banking (ECON4020), the required courses for economics majors.[8]

Another characteristic of the course that I considered was the number of days per week that the course met. This aspect was modeled using two dummy variables, ONEDAY and THREEDAY. As all courses in these data were 3-credit hour courses, this was equivalent to controlling for the length of the class meeting on any given day. For example, 38 percent of upper-level courses met once a week; each meeting was 3 hours in duration. Principles classes were more commonly taught thrice weekly and, less commonly, meet once a week, compared to the upper-division sample. The base category in this case was classes that met twice weekly for $1^1/_2$ hours per lecture.

In many earlier contributions to the literature, researchers have studied the effects of class size on SET scores. Becker and Powers (2001) discussed the sample selection problem inherent in studies such as mine: Students who do not expect to be performing well in class and those who do not like their instructors are more likely to withdraw from a class than are other students. Although the data do not permit a comprehensive treatment of this issue, the findings of Becker and Powers suggest that the appropriate measure of class size is the enrollment at the beginning of the semester because class-size measures that are based on terminal enrollment or an average of initial and terminal enrollment are likely to be endogenous. In the present work, CLSIZE was defined as the number of students enrolled in the class at the beginning of the semester.[9] As class size increases, teaching methods must change. It may be reasonable to assume that students view larger class sizes in a negative manner.[10] The class size in the principles dataset ranged from 19 to 318, with an average of 82.3, whereas the average number of students in upper-level classes was just under 33, with a range from 16 to 53.

I used two dummy variables in an attempt to control for instructor experience. The first had a value of 1 if the instructor in question had between 5 and 10 semesters of experience, and the second took on a value of 1 if the instructor had 11 or more semesters of experience.[11] The base category, then, was courses taught by relatively inexperienced instructors.[12] Thirty-seven percent of principles classes were taught by relatively inexperienced instructors, and 27 percent had very experienced instructors. As one might expect, upper-division classes were more commonly taught by experienced instructors.

To control for the unobservable characteristics of the instructor, course, and semester, and to take advantage of the fact that $8^1/_2$ years of data were available, I used a panel approach, specifically a three-way fixed-effect model.[13] In addition to other explanatory variables, the fixed-effect model I included a dummy variable for each instructor, as well as a dummy variable for each semester. For the present research, the equation of interest was as follows:

$$y_{itj} = \beta_1 + \alpha_i + \gamma_t + \lambda_j + \sum_{k=2}^{K} x_{kitj}\,\beta_k + \varepsilon_{itj}, \tag{1}$$

where α_i represents the fixed-effect specific to instructor i, γ_t represents the fixed-effect specific to semester t, λ_j represents the fixed-effect specific to course j, x_{kitj} includes the class, course, and instructor specific explanatory variables listed above, and ε_{itj} is assumed to be well-behaved.

Isely and Singh (2005) also used a fixed-effect model to examine the determinants of SET scores. However, because their focus was on differences in the way that a given instructor taught different courses, they used a one-way fixed-effect model to examine the variations of SET for a particular instructor, course, and section from the average SET for that instructor in that course as a function of similar deviations of expected grades and other control variables. This was equivalent to having a dummy variable for each course of each instructor. In this arrangement, the fixed-effect coefficient would amount to the intercept for a given instructor teaching a specific course. However, this specification sacrifices the ability to gauge the effect that teaching a particular course may have on SET scores regardless of who the instructor is (that is, factors intrinsic to the particular course but not specific to instructors). Furthermore, the Isely and Singh method did not allow an overall comparison of instructors. One could only say that one instructor rated better than another in a given course.[14]

There are reasons to believe that expected grade is an endogenous variable. Although an instructor's inflating of students' grade expectations might lead to the class assigning better average evaluation scores, if it is true that better teachers receive better evaluation scores, instructors with better evaluation scores will naturally have better performing students who expect higher grades. The empirical evidence on this sort of endogeneity is mixed: Seiver (1983) and Nelson and Lynch (1984) found endogeneity to be a problem, whereas Krautmann and Sander (1999) and Isely and Singh (2005) found the opposite. If EXPGRADE is endogenous, ordinary least squares (OLS) would yield biased and inconsistent parameter estimates. In such cases, these data should be analyzed using a two-stage least squares (2SLS) procedure. I carried out Hausman specification tests for each model to determine whether EXPGRADE was endogenous.

RESULTS

Principles Classes

Because tests for pooling data indicate that it is inappropriate to pool principles and upper-division courses, each subset of the data was considered separately. Hausman specification tests indicated that there was no evidence that EXPGRADE was endogenous, and, as a result, the use of OLS techniques was warranted. The regressions using data only from principles classes are presented in Table 2; there was a significant effect of expected grade on SET scores, with an increase in the average expected grade of 1 point on the usual 4-point scale causing an improvement

TABLE 2. Ordinary Least Squares Regression Results

Variable	Principles classes	Upper-division classes
CONSTANT	2.7962*** (13.572)	2.3189*** (7.276)
EXPGRADE	–0.3417*** (–5.748)	–0.2999*** (–3.556)
PCTMAJ		0.0035** (2.022)
ONEDAY	0.0013 (0.027)	0.0499 (0.805)
THREEDAY	–0.0221 (–0.726)	–0.0017 (–0.027)
CLSIZE	0.0008*** (2.703)	0.0012 (0.467)
EXPERIENCE: 5–10 SEMESTERS	–0.1002** (–2.117)	0.0301 (0.276)
EXPERIENCE: 11+ SEMESTERS	–0.1728* (–1.836)	0.1975 (1.011)
RESPONSE	0.0013 (1.262)	0.0029** (1.996)
Course Fixed Effects		
ECON1100: Principles of micro	–0.0359 (–1.148)	
ECON3000: Contemp. issues		–0.1633 (–0.754)
ECON3050: Consumer		–0.0802 (–0.354)
ECON3150: Discrimination		–0.1641 (–1.422)
ECON3560: Macro		0.0562 (0.412)
ECON4020: Money and banking		–0.0661 (–0.408)
ECON4100: Comparative systems		0.0885 (0.281)
ECON4140: Managerial		0.0614 (0.353)
ECON4150: Public finance		–0.1111 (–0.968)
ECON4180: Health		–0.0745 (–0.447)
ECON4290: Labor		–0.5279*** (–3.241)
ECON4440: Environmental		0.0583 (0.369)
ECON4460: Ind. organization		0.0026 (0.010)
ECON4500: Sports		–0.4618** (–2.116)
ECON4510: History of thought		–0.1251 (–0.830)
ECON4600: Development		–0.1149 (–0.683)
ECON4630: Research methods		–0.3671 (–1.393)
ECON4850: Trade		–0.2209* (–1.742)
ECON4870: Econometrics		–0.3513 (–1.077)
Sample size	360	247
$\bar{R}^2$	0.779	0.646
F statistic	25.360	8.230

Notes: t statistics are in parentheses. *significant at a two-tailed Type I error level of .10. **significant at a two-tailed Type I error level of .05. ***significant at a two-tailed Type I error level of .01. Dependent variable is measured on a scale of 1 to 4, with lower scores representing better teaching evaluation scores. Estimates of the instructor- and semester-specific fixed effects are available from the author upon request.

in SET scores of about 0.34 points. The implication was that at the introductory level better teaching evaluation scores can be bought by instructors causing students to expect higher grades. The magnitude of the coefficient was comparable to that reported by Isely and Singh (2005) in their study of classes at Grand Valley State University but smaller than Dilts (1980) found at Ball State University or Krautmann and Sander (1999) at DePaul University.

Several characteristics of particular classes were important determinants of evaluation scores.[15] The number of days per week the class met was not an important determinant of SET scores in either a statistical or an economic sense. This finding was somewhat surprising given that several earlier studies (Nichols and Soper 1972; Nelson and Lynch 1984; Isely and Singh 2005) found such effects to be important. This difference may be indicative of heterogeneity across universities or that these earlier studies failed to account for the several sources of unobserved heterogeneity considered here. Class size had a significant effect on SET scores; a one-student increase in class size caused evaluation scores to rise (worsen) by 0.0008 points. To understand this finding, consider two classes that are identical except that the first is average in terms of class size (82 students), and the second is one standard deviation greater than the mean (127 students). The evaluation score for the former would be 0.036 points better than the latter. The magnitude of this effect is similar to that found by Danielsen and White (1976) using data from the University of Georgia but smaller than that found by Krautmann and Sander (1999) and by Isely and Singh (2005). This improvement would have some effect on the rankings of instructors and is evidence that smaller class sizes are to be preferred in this regard.

Experience was of considerable importance in determining SET scores. In particular, instructors with between 5 and 10 semesters of experience had SET scores that were about 0.10 points better than instructors with less than 5 years' experience, *ceteris paribus*. Similarly, instructors with 11 or more semesters of experience had SET scores that were 0.17 points better than their less-experienced colleagues. The response rate was not a significant determinant of SET scores.

Fixed effects were generally important, with instructor-specific fixed effects especially so.[16] In particular, there was no significant difference between principles of microeconomics and principles of macroeconomics sections. However, the majority of the instructor-specific effects were different from zero in a statistical sense and were large in magnitude. As discussed later, it may be useful to consider these coefficients to be longer term measures of teaching quality, and, as such, instructors could be ranked accordingly. The best score in this regard belonged to instructor 3, and the highest (worst) score was associated with instructor 18. An *F* test of the null hypothesis that the instructor-specific fixed effects were jointly zero can be rejected at the 99 percent confidence level.[17] Only three of the semester-specific fixed-effects coefficients were statistically significantly different from zero. Nevertheless, an *F* test of the null hypothesis that the semester-specific fixed-effects were jointly zero can be rejected at a 99 percent confidence level.[18]

The regression that used the alternative dependent variable noted above produced very similar results, with only one notable difference. The estimated coefficient on the dummy representing 11 or more years of experience, although similar in magnitude to the regression reported in Table 2, was not statistically significant.

Upper-Division Classes

As was the case with the principles regressions, Hausman specification tests for the upper-division courses indicated that EXPGRADE was not endogenous, and so, once again, a simple one-stage FEM would be the appropriate specification.

First, the data allowed us to reject the hypothesis that the coefficient on EXPGRADE was zero (Table 2). The magnitude of the coefficient was comparable to that in the principles regression. This result was in opposition to Seiver (1983) and Nelson and Lynch (1984) but in accord with Isely and Singh (2005). Evidently, there was a significant negative relationship between EXPGRADE and SET score, implying that instructors might be able to increase their evaluation scores by inflating grade expectations.

Classes containing high proportions of economics majors seemed to be more critical of instructors than other classes. An increase in the percentage of the class that was majoring in economics worsened SETs scores by about 0.004 points. Instructors of a class made up exclusively of economics majors would receive evaluation scores that were about 0.20 points worse than a teacher of an identical class in which only half were majors. This result was somewhat surprising, because one might have reasonably expected nonmajors to be less appreciative of instruction in economics classes. It may be the case that nonmajors were more likely to have elected to take the class out of interest, whereas economics majors were more likely to have been required to take a given economics class. As was the case with principles classes, the number of days per week that an upper-division class met had no evident effect on SET scores once other factors were considered.

It is interesting to note that upper-division class size had no apparent effect on evaluation scores, unlike the situation with principles classes. This might be because there was much less variation in class sizes in upper-division courses, which ranged in size from 16 to 53 students, whereas principles classes ranged from as small as 19 to as large as 318. Unlike the principles case, instructor experience in upper-division teaching did not affect SET scores. Because the vast majority of upper-division classes were taught by faculty members holding doctoral degrees, students in a given class might perceive their instructor to be an expert regardless of the level of the instructor's experience.

The response rate was a statistically significant determinant of SET scores. Classes in which a higher proportion of students participated in the evaluation process tended to be more critical of their instructors, with each additional percentage point of attendance associated with a worsening of SET scores of around 0.003 points. Although this effect had statistical significance, it was rather small in magnitude.

Table 2 also reveals the extent to which the course fixed-effects were significant determinants of SET score. SET scores were significantly different from the base category (intermediate microeconomics) for only a few particular upper-division classes, but the coefficients for these were rather large. For example, teachers of the labor economics class (ECON4290) had SET scores approximately 0.53 points better than instructors of intermediate microeconomics. An effect of a similar magnitude existed for the sports economics (ECON4500) class and, to a lesser extent, international trade (ECON4850) and research methods (ECON4630).[19] Twenty different instructors taught upper-division courses during the 17 semesters spanned by the data. As was the case with the principles regressions, the majority of the instructor-specific dummies were significantly different from zero and large in magnitude, indicating that factors peculiar to instructors form an important part of SET scores. These coefficients ranged in magnitude from about 0.46 to –0.37.[20] Finally, several semester-specific fixed effects were significantly different from zero.[21] Together, these findings suggested that much of what students considered when evaluating teachers involved difficult-to-measure aspects of the instructor, and to a lesser extent, the semester in which the course was taught and the course itself.

Once again, the specification that employs the alternative dependent variable produced very similar results to those reported in Table 2. The only important differences involved the statistical significance of the course-specific fixed-effects. In the alternative specification, teaching economics of discrimination (ECON3150) significantly improved an instructor's SET score, whereas teaching international trade (ECON4850) did not have a statistically significant effect.

Grade Surprise: Evaluating the Isely and Singh Variable

Before leaving this topic, it is useful to compare further my results with those of Isely and Singh (2005). Despite the concerns already noted about the variable, I constructed a grade surprise variable similar to that suggested by Isely and Singh. Following Isely and Singh, I carried out a series of Davidson-MacKinnon (1981) *J* tests to determine whether the specifications using EXPGRADE (model I) were superior to those employing the grade surprise variable (model II). The first step of the test involved estimating model II and calculating the fitted values from that regression. Model I was then estimated, with the fitted values from the first regression included among the regressors in the second regression. If the coefficient on this fitted value was found to be significantly different from zero, the implication was that model I was not the correct specification (otherwise it was). The second step was to reverse this procedure, estimating model I in the first step. Should the estimated coefficient on the fitted value be significantly different from zero, model II was the preferred specification. Logically, this means that the Davidson-MacKinnon *J* test might indicate that model I was clearly superior to model II, that model I was clearly inferior to model II, or that it was inconclusive. This latter finding would emerge if the coefficients on the fitted values from both parts of the test were found to be either significantly different from zero or both not significantly different from zero.

For the principles regressions, the Davidson-MacKinnon *J* test was inconclusive. The *t* statistic from the first step was 2.216; this implied that model I (the model with EXPGRADE) was not the correct specification. However, the *t* statistic from step two was 6.190, implying that model II (the model with the grade surprise variable) was also not the preferred specification.[22]

The Davidson-MacKinnon *J* test involving the upper-division classes indicate that the model that used EXPGRADE was preferable to that using the Isely and Singh variable. The *t* statistics for the first step was –0.859, implying that model I was the "true" model. The *t* statistic for the second step was 3.488, implying that the grade surprise specification was not preferred.

In short, for upper-division classes, Davidson-MacKinnon *J* tests indicated that specifications using EXPGRADE were preferred to those using the Isely and Singh variable. For principles classes, the surprise variable was not obviously preferable. For this reason, as well as because of the serious concerns noted earlier regarding the manner in which a grade surprise variable was calculated, EXPGRADE was used in all specifications in this article.[23]

ADJUSTING RANKINGS OF INSTRUCTORS

Several researchers in this area (Danielsen and White 1976; Mason, Steagall, and Fabritius 1995) have suggested adjusting raw SET scores to eliminate the influence of factors that either could be manipulated by instructors to their advantage (e.g., expected grade) or that might be beyond an instructor's control (such as type of course). The model presented above suggests at least two adjustments: a ranking based on the magnitude of the estimated fixed-effects coefficients, and a ranking based on an adjustment of each semester's raw SET score that accounts for extrinsic influences.

Fixed-effect Rankings

Instructors could be ranked according to their fixed-effect coefficients. In essence, an instructor's coefficient is the amount by which his or her intercept varies from the overall intercept that is common to all instructors. For example, for instructors of principles classes, the smallest and largest instructor-specific coefficients were –0.450 and 0.991, respectively (Table 3). Given the overall constant of 2.796 and a semester-specific fixed-effect of –0.207 in the fall 1994 semester, this implied that the fall 1994 intercept for the first instructor was 2.139, and for the second was 3.580. A comparison of these numbers held constant all observable effects as well as time-specific effects. It may be appropriate to think of these numbers as longer term measures of instructor quality. For example, the instructor 22's average SET score over all semesters in principles classes was 1.674. Out of the 28 principles instructors, this particular instructor would rank as seventh best. However, when ranked according to the fixed-effect coefficient, instructor 22 falls to the 12th position.

The rankings based on average SET score and the fixed-effect coefficients were relatively highly correlated, with Spearman's rank correlation coefficients equal to at least 0.95 for principles classes and at least 0.88 for upper-division classes (both were significantly different from zero in a statistical sense). This reflected the fact that the use of the fixed-effect coefficient changed an instructor's ranking by two or fewer positions in about half the cases. Still, certain instructors would be affected by the use of ranking based on the fixed-effect coefficients in a dramatic fashion. As previously noted, among principles instructors, instructor 22 was an example of a person who would see his or her ranking fall dramatically if fixed-effect coefficients were used to construct rankings. Instructor 44 was an example of a principles

TABLE 3. Comparison of Rankings Based on Average SET Score and Fixed-Effect Coefficients

	Principles instructors				Upper-division instructors			
Instructor	Average of EVAL (st. dev.)	Ranking based on average of EVAL	Fixed-effect coefficient (st. error)	Ranking based on fixed-effect coefficient	Average of EVAL (st. dev.)	Ranking based on average of EVAL	Fixed-effect coefficient (st. error)	Ranking based on fixed-effect coefficient
α_3	1.447 (0.111)	1	–0.450 (0.082)	1				
α_8	2.222 (0.269)	22	0.354 (0.031)	24				
α_7					1.989 (0.126)	17	0.462 (0.162)	19
α_9	1.518 (0.076)	2	–0.287 (0.040)	5	1.334 (0.158)	2	–0.133 (0.208)	8
α_{11}	1.831 (0.090)	18	0.003 (0.049)	18	1.901 (0.079)	15	–.016 (0.127)	12
α_{14}	1.772 (0.252)	15	–0.052 (0.051)	16	1.477 (0.323)	8	–0.120 (0.120)	9
α_{15}	1.547 (n.a.)	4	–0.356 (0.122)	2				
α_{16}	2.261 (0.047)	24	0.349 (0.099)	23				
α_{18}	2.706 (0.237)	27	0.991 (0.151)	28	2.224 (0.196)	19	0.321 (0.102)	17
α_{20}	2.048 (n.a.)	19	0.184 (0.121)	21				
α_{21}					1.646 (0.124)	12	0.029 (0.153)	13
α_{22}	1.674 (0.138)	7	–0.108 (0.115)	12	1.460 (0.247)	5	–0.290 (0.091)	3
α_{23}	2.826 (0.401)	28	0.855 (0.071)	27				
α_{24}	1.645 (0.162)	6	–0.309 (0.062)	4	1.543 (0.125)	9	–0.374 (0.273)	1
α_{25}	1.751 (0.171)	11	–0.136 (0.069)	9				
α_{26}	1.753 (0.084)	14	–0.127 (0.064)	11				
α_{27}					1.889 (0.343)	14	0.141 (0.078)	15

α_{28}	2.600 (0.409)	26	0.648 (0.095)	26				
α_{31}	1.689 (0.127)	8	–0.127 (0.097)	10				
α_{32}	2.111 (0.203)	21	0.148 (0.061)	20				
α_{33}	1.716 (n.a.)	10	–0.189 (0.106)	7				
α_{36}	1.813 (n.a.)	16	–0.076 (0.144)	14	1.441 (0.350)	4	–0.215 (0.261)	4
α_{37}	1.751 (0.208)	12	–0.027 (0.080)	17	2.057 (0.340)	18	0.091 (0.122)	14
α_{38}					1.966 (0.339)	16	0.525 (0.202)	20
α_{41}	2.061 (0.263)	20	0.105 (0.107)	19				
α_{42}					1.414 (0.174)	3	–0.056 (0.133)	10
α_{44}	1.752 (0.083)	13	–0.189 (0.090)	8				
α_{47}	1.713 (0.148)	9	–0.094 (0.036)	13	1.595 (0.087)	11	–0.054 (0.123)	11
α_{49}	2.333 (0.285)	25	0.435 (0.064)	25				
α_{50}	2.243 (0.289)	23	0.330 (0.116)	22	2.239 (0.313)	20	0.395 (0.120)	18
α_{52}					1.472 (0.202)	7	–0.206 (0.174)	5
α_{53}	1.534 (0.146)	3	–0.277 (0.045)	6	1.309 (0.106)	1	–0.342 (0.169)	2
α_{54}	1.586 (0.080)	5	–0.332 (0.090)	3	1.470 (0.076)	6	–0.156 (0.131)	7
α_{55}	1.829 (0.182)	17	–0.074 (0.061)	15	1.586 (0.340)	10	–0.185 (0.226)	6
α_{61}					1.651 (0.349)	13	0.213 (0.217)	16
Pearson's ρ		0.953				0.881		
t statistic		16.127				8.552		

instructor who would presumably prefer the fixed-effect rankings. For upper-division classes, instructor 24 jumped from 9th place (out of 20) in the average SET score, ranking to the best score, if fixed-effect coefficients rankings were used.

In short, these rankings were similar, but there were a few substantial differences. Ranking instructors by their fixed-effect coefficients may give a more complete picture of relative teaching quality in the sense that it takes into account factors that may be beyond an instructor's control.

Semester-by-Semester Rankings

If the principal interest were in adjusting scores for each semester, a comparison of the fixed-effect coefficients would not serve. There is one fixed-effect coefficient for each instructor based on information from all semesters. Most universities evaluate faculty each semester, and a given semester's performance might be better or worse than the overall trend for that instructor. As an alternative, suppose that for a particular semester, I start with the raw SET score but remove the influence of the observable extrinsic influences. For example, the results above suggest that the instructor can leverage better SET scores by causing students to expect higher grades. An alternative ranking could remove the rewards for such behavior. Other variables cause an instructor's evaluation score to worsen (for example, teaching an upper-division class with a large percentage of economics majors); it would be useful to compensate for such penalties. Mathematically, this adjustment could be represented as follows:

$$\tilde{y}_{itj} = y_{itj} - \hat{\lambda}_j - \sum_{k=2}^{K} x_{kitj}\,\hat{\beta}_k, \tag{2}$$

where $\tilde{y}_{itj}$ is the adjusted SET score, y_{itj} is the official SET score, $\hat{\lambda}_j$ is the estimated fixed-effect for course j, and $\hat{\beta}_k$ represents estimated parameters from equation (1). More experienced instructors should be allowed to receive whatever benefit accrues to them in the form of better evaluation scores. For this reason, the experience dummies were not used in the adjustment in the regressions. The adjustment did take into account the effects of expected grade, number of days per week the class meets, class size, response rate, and (in the case of upper-division classes) percentage of the class that was majoring in economics. In essence, equation (2) produces a ranking stripped of factors that can be manipulated by instructors to their advantage[24] and other factors beyond instructors' control but allows experience and otherwise unobservable instructor-specific effects to remain.[25]

Rankings based on the official or raw SET scores and the rankings based on the adjustment were generally relatively highly correlated, with Pearson's rank-correlation coefficients for principles classes ranging from 0.736 to 0.982 and for upper-division classes, from 0.709 to 0.964. With few exceptions, the official ranking and the adjusted ranking for any particular semester had a Pearson's correlation coefficient of at least 0.8.

Despite the high degree of correlation, important differences were caused by the adjustment of the rankings. First, in most semesters, there were faculty members who moved up or down several positions in the rankings after adjustment, although in many cases, the adjustment caused movement of only one position in the rankings, if any. Instructor 24's rankings in principles classes in the spring of 1998 illustrates this point. This individual was rated as the seventh best teacher of principles classes out of 11 instructors if the rankings based on raw SET scores were employed. His or her position rose to the top spot if the rankings were adjusted. When considering instructors of upper-division classes, the adjustment once again worked to the benefit of instructor 24 in the spring 1995 semester—he or she moved from fifth to first place out of 11. Other instructors were affected (for better or worse) by this adjustment. Even a movement of one position could have important implications for a particular faculty member in personnel decisions, the allocation of merit-raise money, and the general esteem of colleagues.

A second and related point involves whether, over the course of many semesters, a particular faculty member is consistently over- or under-valued by the official rankings. If an instructor's official SET score ranking is either consistently above or consistently below his or her corrected ranking, over time some inequities might develop. To explore this possibility, I averaged the rank based on the raw SET score of a given instructor over all semesters in which he or she had taught and compared that with that instructor's average rank based on the corrected scores. In most cases, the correction did not have a large effect on an instructor's average ranking. That is, in most cases, an instructor might see his or her ranking change in the instructor's favor in one semester, but to his or her detriment in other semesters, largely balancing out over time. However, this was not the case for every instructor. For example, of the 12 semesters that instructor 24 has taught principles classes, in 8, the adjusted rankings were in the instructor's favor, and in 3, the adjusted ranking represented no change from the ranking based on the official SET score. In only one semester would this instructor's position in the rankings be adversely affected by the proposed adjustment. If rankings were adjusted each semester, instructor 24's average rank would be nearly two positions higher than is the case at present. The differences are even more important for faculty members teaching upper-division classes. Instructor 14, for example, would see his or her average ranking fall from 4.4 under the official ranking to 6th best under the adjustment. Presumably, over the years covered by these data, anyone evaluating faculty teaching based on an instructor's SET scores would have fairly consistently undervalued instructor 24's contribution and would have overvalued that of instructor 14. Quite a number of other instructors would be affected in a similar manner.

CONCLUSIONS

Even in the unlikely event that SET scores contain no information about the quality and effectiveness of teaching, the fact remains that they are widely used by instructors and administrators to evaluate teaching. This alone makes a better understanding of the determinants of SET scores worth pursuing.

Efforts to isolate the variables explaining SET scores have been made for more than 40 years. Unfortunately, early efforts suffered from one or more serious shortcomings in the statistical methods used, and all research has been hampered to some extent by the unavailability of data from more than two or three consecutive semesters. This research is an effort to correct the previous problems and examine more completely the determinants of SET scores using a fixed-effect model. This specification deals appropriately with the unobserved course-specific, instructor-specific, and semester-specific effects that may affect SET scores. I also tested for endogeneity of expected grade. The data cover $8^1/_2$ academic years and so offer a unique glimpse into how the passage of time might affect SET scores.

Statistical tests indicate that it is inappropriate to pool principles and upper-division classes when examining SET scores, a finding that future research should take into account. A principal empirical finding involves the evidence regarding the possible contamination of SET scores by instructors attempting to buy better SET scores by raising grade expectations. In particular, higher expected grades do lead to significantly better SET scores among both principles and upper-division classes. This issue has been hotly debated in the literature, and this debate will surely continue. In any case, this finding underlines the importance of adjusting instructor rankings to remove any such effect.

The results of the present research also demonstrate that class size may affect SET scores, at least at the principles level. This finding indicates that teaching smaller classes results in better SET scores, *ceteris paribus*. If it is true that better SET scores are correlated with measures of student learning, this result reinforces the commonly held view that teaching is most effective in relatively small class sizes. In addition, certain other student and course attributes, such as the percentage of the class that is majoring in economics and the response rate, significantly influence SET scores in upper-division classes.

Unobserved course-specific effects are important determinants of SET scores in upper-division classes, with instructors of labor economics, sports economics, and international trade receiving better evaluation score than their colleagues, *ceteris paribus*. It is perhaps not surprising that there are no distinctions in principles classes between SET scores of instructors teaching microeconomics and those teaching macroeconomics.

Experience of instructors seems to have an important relationship with SET scores in principles classes, although it appears to be unimportant in upper-division classes. It is also important to note that the unobservable characteristics of instructors and to a lesser extent semesters (as captured by the fixed-effect coefficients) are typically large in magnitude and statistically significant. That is, these unobservable effects have at least as strong an influence on a typical instructor's SET scores as all other effects combined.

Adjusting rankings of instructors to account for influences beyond their control yields rank orderings that are relatively highly correlated with rankings based on raw or official SET scores. Nevertheless, important differences exist for certain faculty members. Given that many colleges and universities use rankings of instructors by SET scores as factors in promotion and tenure decisions, other personnel decisions, and the allocation of merit raises, the results presented here suggest that rankings adjustments may be appropriate and long overdue.

NOTES

1. An extensive review of this literature is available from the author on request.
2. Department of the Economics at UNT is part of the College of Arts and Sciences. However, many students from the College of Business take economics classes, and a one-degree program is jointly administered by the two colleges.
3. The value of the *F* statistic in this case is 2.00, and because the critical value of the test at the 95 percent confidence level is 1.75, I rejected the null hypothesis that pooling is appropriate.
4. It may not be appropriate to pool upper-division classes. Data constraints prevented this issue from being tested, but I presume that the inclusion of course-specific dummy variables dramatically lessens this potential problem.
5. Siegfried and Vahaly (1975) presented evidence that announcing evaluations in advance does not introduce any particular bias.
6. Because the second and third questions are phrased in a negative manner, these were rescaled by subtracting each from 5. Each question received equal weight.
7. The gender composition of the class was also considered, but the percentage of the class that was female was not a significant determinant in any specification.
8. The effect that the time of day that a given class meets might have on SET scores was also considered but, in no case, were these effects significant in a statistical or economic sense.
9. Kennedy and Siegfried (1997) noted that class size could also be endogenous if better teachers were assigned to larger classes, but they found no evidence of this.
10. It is possible that the relationship between class size and SET score is nonlinear: Perhaps above a certain class size further increases in the number of students is perceived in a positive light by students because of advantages of anonymity. To examine this possibility, I also included the square and cube of class size as regressors in the models presented here. In all cases, I failed to reject the hypothesis that the class size-SET score relationship is a linear one.
11. The data do not include information about semesters of experience teaching prior to an instructor joining the faculty at UNT.
12. Neither the number of classes a given instructor taught in a given semester nor whether the instructor had recently taught a particular course were significant determinants of SET scores in any specification and therefore were not included in the models.
13. In the models using principles classes, the relatively large Hausman statistics for the models examined argue for the use of the fixed-effects (FEM) rather than the random-effects (REM) framework. The *p* values for these statistics were in both cases below 0.05. However, the Hausman statistics were smaller for the models using upper-division classes, meaning that one cannot reject the hypothesis that the REM is appropriate. As it happens, results from the FEM and the REM were essentially identical, and because the use of the FEM makes adjusting instructor rankings substantially simpler, the FEM was used. REM results are available from the author upon request.
14. To explore how the differences in specification might affect the results, I applied the Isely and Singh (2005) approach to the data I used in this research. My specification assumed that the influence of instructor and the influence of course are additive and separate. In fact, the specification used here was a restricted version of that used by Isely and Singh, so an *F* test of the hypothesis that it is appropriate to treat the fixed effects as I did can be carried out. The *F* statistics were 0.41 and 0.79 for the principles and upper-division data, respectively, so the assumption that separate and additive effects are appropriate cannot be rejected. Furthermore, there were no important differences in the results, either in the magnitudes of the estimated nonfixed-effect coefficients or in their estimated standard errors. These results are available from the author upon request.
15. The percentage of the class that was majoring in economics was not included as a regressor in the principles regressions because less than 1 percent of students in these classes were economics majors.
16. In the results that follow, I present an overall constant that was recovered in the manner described by Greene (2000, 565). Each instructor-specific fixed-effect represents by how much that instructor's intercept differed from the overall constant for any particular time period. Similarly, each semester-specific fixed-effect represents the amount by which a particular semester's intercept differed from the overall constant for any particular instructor.
17. The *F* statistic was 21.43.
18. The *F* statistic was 2.35. The critical value in this case was 2.02.
19. An *F* test of the hypothesis that the course-specific fixed-effects are jointly insignificant can only be rejected at the 0.20 Type I error level (the *F* statistic was 1.28). Nevertheless, these dummies were included on the argument that it is important to control for course-specific heterogeneity.
20. With an *F* statistic of 4.65, the hypothesis that the instructor-specific effects jointly did not matter can be rejected at the 0.01 Type I error level.

21. The hypothesis that semester-specific effects jointly did not matter can be rejected at the 0.12 Type I error level (the *F* statistic equals 1.44).
22. I also examine the Isely and Singh (2005) specification that used fixed effects for each course of each instructor (see note 14). I also applied a Davidson-MacKinnon *J* test to this specification, but the conclusion was the same: The surprise variable was not clearly preferable to EXPGRADE.
23. It should be noted that the larger class sizes and lower response rates in the present data compared with that used by Isely and Singh may make it more likely that specifications that employ the surprise variable will be rejected.
24. Expected grade, which comes from the evaluation process, should be used in the adjustment rather than actual realized grade. If the latter were used, instructors could "game" the process by leading students to expect high grade but then assigning them low ones. The use of expected grade removes this possibility.
25. Semester-specific effects were not included in the adjustment proposed in equation (2), on the grounds that they affect each instructor in a given semester equally.

REFERENCES

Aigner, D. J., and F. D. Thum. 1986. On student evaluation of teaching ability. *Journal of Economic Education* 17 (Fall): 243–65.

Becker, W. E., and J. Powers. 2001. Student performance, attrition, and class size given missing student data. *Economics of Education Review* 20 (August): 377–88.

Becker, W. E., and M. Watts. 1999. How departments of economics evaluate teaching. *American Economic Review Papers and Proceedings* 89 (2): 344–49.

Danielsen, A. L., and R. A. White. 1976. Some evidence on the variables associated with student evaluations of teachers. *Journal of Economic Education* 7 (Spring): 117–19.

Davidson, R., and J. G. MacKinnon. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49 (3): 781–93.

Dilts, D. A. 1980. A statistical interpretation of student evaluation feedback. *Journal of Economic Education* 11 (Spring): 10–15.

Greene, W. H. 2000. *Econometric analysis*. 4th ed. Upper Saddle River, NJ: Prentice-Hall.

Heilman, J. D., and W. D. Armentrout. 1936. Are student ratings of teachers affected by grades? *Journal of Educational Psychology* 27 (March): 197–216.

Isely, P., and H. Singh. 2005. Do higher grades lead to favorable student evaluations? *Journal of Economic Education* 36 (Winter): 29–42.

Kennedy, P. E., and J. J. Siegfried. 1997. Class size and achievement in introductory economics: Evidence from the TUCE III data. *Economics of Education Review* 16 (4): 385–94.

Krautmann, A. C., and W. Sander. 1999. Grades and student evaluations of teachers. *Economics of Education Review* 18 (1): 59–63.

Mason, P. M., J. W. Steagall, and M. M. Fabritius. 1995. Student evaluations of faculty: A new procedure for using aggregate measures of performance. *Economics of Education Review* 14 (4): 403–16.

Nelson, J. P., and K. A. Lynch. 1984. Grade inflation, real income, simultaneity, and teaching evaluations. *Journal of Economic Education* 15 (Winter): 21–37.

Nichols, A., and J. C. Soper. 1972. Economic man in the classroom. *Journal of Political Economy* 80 (5): 1069–73.

Seiver, D. A. 1983. Evaluations and grades: A simultaneous framework. *Journal of Economic Education* 14 (Summer): 32–38.

Siegfried, J. J., and J. Vahaly, Jr. 1975. Sample bias of unannounced student evaluations of teaching. *Journal of Economic Education* 6 (Spring): 137–39.

Tronetti, R. J., 2001. Does class size matter? Evidence from panel data estimation. Master's thesis, University of Central Florida.

Do Higher Grades Lead to Favorable Student Evaluations?

Paul Isely and Harinder Singh

Abstract: The relationship between expected grades and student evaluations of teaching (SET) has been controversial. The authors take another look at the controversy by employing class-specific observations and controlling for time-invariant instructor and course differences with a fixed-effects model. The authors' empirical results indicate that if an instructor of a particular course has some classes in which students expect higher grades, a more favorable average SET is obtained in these classes. Moreover, they find that it is the gap between expected grade and cumulative grade point average of incoming students that is the relevant explanatory variable, not expected grade as employed in the previous literature.

Key words: expected grades, fixed-effects model, grade point average (GPA), relative grade level, student evaluations of teaching (SET)
JEL code: A22

A positive gross correlation between student evaluations of teaching (SET) and student grades has been found in previous empirical work. The positive correlation could result from a variety of factors. Given the fact that SET are employed quite extensively for faculty evaluations procedures, it is conceivable on *a priori* grounds that some instructors at different times may grade more leniently to obtain more favorable SET. Students could infer their own ability or course quality from higher grades. Consequently, they may reward instructors with higher SET. A bias could result from student self-selection. The instructor's reputation could attract high achievers or low-performing students may be induced to drop the class. SET and grades may be influenced by other poorly observed instructor specific variables such as an outgoing personality, laid-back style, and other forms of behavior that are difficult to measure directly.

Given a variety of intangible factors that influence teaching, it is not surprising that although most investigations have found a positive relationship between SET and student grades, the empirical evidence is somewhat mixed. Nelson and Lynch (1984) found that favorable student grades improve SET by 0.15 in ordinary least

Paul Isely *is an associate professor of economics (e-mail: iselyp@gvsu.edu), and* ***Harinder Singh*** *is a professor of economics at Grand Valley State University.* *The authors are grateful to Paul Sicilian, Gerald Simons, and Paul Thorsnes for valuable comments, and to Thelma Cathey, Ramey DeGama, and Madeleine Moore for research assistance, and to the Pew Teaching and Learning Center for a grant. They benefited enormously by feedback from Peter Kennedy and several anonymous referees.*

squares (OLS) estimates on a 4.0 scale. However, they found the relationship to be statistically insignificant in a simultaneous equations model (SEM). Seiver (1983), although making a case for employing a simultaneous framework, found a significant relationship in the OLS model but not in the SEM. DeCanio (1986), employing a multinomial logit approach, found no significant relationship between expected grades and SET. The literature in psychology journals about this issue is equivocal as well. A good flavor of this controversy in psychology is provided in a series of review articles in the November 1997 issue of the *American Psychologist* (Greenwald 1997).

In general, these studies rely on ad hoc specifications that control for a variety of characteristics of students, courses, and instructors. Their coefficient estimates come from variation in SET that may result from different courses, different instructors, and a different student draw. A major problem with such studies is that unmeasured characteristics of courses and instructors can create bias. What we really wanted to know was the answer to the following question: If the *same* professor teaches the *same* course several times, but students in some classes (of the same course) expect higher grades, will that expectation lead to more favorable SET scores? One contribution of this article is to address this question explicitly, rather than indirectly as is done in the literature, by using a fixed-effects model. By estimating instructor and course-specific fixed effects in all our specifications, we controlled for the intangible time-invariant variables that may bias existing estimates. We found that higher expected grades do influence SET.

Contrary to previous literature that has employed *expected grade* as an explanatory variable in empirical specifications, we found that it is the *difference between expected grade and cumulative GPA* (grade point average) that affects SET more significantly. The mixed results in the previous literature may be attributed to the fact that the gap between expected grade and cumulative GPA is the preferred explanatory variable. In addition, our results did not appear to be influenced by a simultaneous equation bias.

METHODS

The variation in SET can be attributed to three major sources: differences in (a) courses, (b) students, and (c) instructors. We controlled for differences in courses by a variety of class-specific variables. Variation in students was controlled by variables such as percentage of students taking a required or optional course, whether the course was in their major or an elective, and the cumulative average GPA of incoming students. Time-invariant instructor and course differences were controlled by a fixed-effects model.

Fixed-Effects Approach

Watts and Bosshardt (1991) showed that instructor-related differences can be captured by a fixed-effects model. They pointed out that “instructors use many different approaches, according to their own interests, attributes, and costs” (336). Other differences may include instructor's gender, charisma, teaching ability, demeanor and deportment, and so forth. On the basis of these considerations, we tested the positive association between GPA and SET using a fixed-effects model that controlled for time-invariant instructor characteristics.

However, the same professor could obtain a significantly different SET depending on which courses he or she teaches. Our course-specific control variables may not incorporate all factors that lead to a variation in SET scores across classes. Some specific courses may be more susceptible to higher or lower SET scores. In all our specifications, the fixed-effects model controlled not only for instructor-related effects but also for course-related effects. We asked the question, If the *same* professor teaches the *same* course several times, but students in some classes (of the same course) expect higher grades, will that expectation lead to more favorable SET scores? By estimating instructor and course specific fixed effects in all our specifications, we controlled for the intangible time-invariant variables that may bias existing estimates.

Class-Specific Data

We employed class-specific, rather than student-specific, data, placing the emphasis on data actually used for faculty evaluation. Nichols and Soper (1972), employing class-specific data, found that higher expected classroom grades improved student evaluations by 0.53 (on a 4.0 scale). Marsh (1987) argued that the appropriate unit of analysis should be class-average SET and grades. There are several reasons why an approach based on class-specific data may provide useful insights.

First, because most applications of SET are based on class averages (for instance, deans look at average SET scores), analyzing the average SET may be useful. Although some *individual* students might reward higher expected grades with favorable evaluations of teachers, the overall mean SET and mean expected grade could have a different relationship. Second, in most previous studies, regression analysis was performed on a set of variables that was a mixture of student-specific (expected grade and SET) and course-specific variables (level of course, whether the class is required, location and time of class, gender of instructor, size of class, etc.). In our analysis, all of the variables were class specific to ensure consistency across observations. Third, our dependent variable was average SET in a specific class (a continuous variable); therefore, we bypassed the problems associated with a dichotomous variable.

One disadvantage of class-specific observations is that one cannot control directly for student characteristics. No direct measures of student learning (such as a pre- and posttest differential) were available to us. We had information about some characteristics of the students such as percentage of students taking a required or optional course and whether the course was in their major or an elective. We also included the cumulative average GPA of the students in each class as a control variable. Assuming a random sample of students, as the number of students incorporated in the study increased, the effect of individual variations was likely to be less important. Given our average class size of 35 students and 260 class observations for our large sample, approximately 9,100 student responses were indirectly incorporated in our analysis.

Gillmore and Greenwald (1999) argued that controlling for the level of difficulty in a course is important for analyzing the impact of expected GPA on SET. We used two proxies to capture the level of difficulty in each class: How students rated the class on a difficulty scale and whether they believed the class was taught too fast. Becker and Watts (1999) pointed out, "to date, correlation studies have not adjusted for the sample selection associated with student withdrawals, or with absenteeism on the day evaluations are administered" (344). Becker and Powers (2001) have also argued that the class size at the beginning of the semester has a more significant impact, especially in large sections. Consequently, we used the class size at the beginning of the semester and included the proportion of students who had withdrawn and those not responding to the SET in each course at the end of the semester as a control variable.

DATA SPECIFICS

We need to address several issues about the data sample, specification of the dependent variable, and the control variables.

Data Sample

Our data sample for economics and finance courses consisted of 260 observations (179 classes in economics and 81 classes in finance). Approximately 50 of the classes were at the principles level. In addition, our data set included a wide array of upper-division electives from Grand Valley State University (GVSU), a state university in Michigan. Classes are taught primarily at the Allendale main campus, downtown Grand Rapids (campus 2), and at nearby Holland (campus 3). Our sample did not include graduate courses. The data were pooled over five academic years, 1994–99. The average panel size for each instructor was 22 classes and 6 classes for an instructor teaching the *same* course.

Faculty members in the Seidman School of Business at GVSU are evaluated annually for salary increases based on performance criteria that include teaching, research, and service. GVSU is a typical state school that places a major emphasis on teaching. The American Assembly of Collegiate Schools of Business (AACSB) accredits all courses at Seidman. Approximately 50 percent of the overall performance criteria are based on SET. Consequently, instructors are concerned about the SET scores they obtain. The average SET score for each course can range from one (*excellent*) to five (*poor*).

Model Specification

The SET survey instrument is provided in the appendix. Our dependent variable was the average SET score for all questions that relate to teaching effectiveness (question numbers 1 to 25). A response of *strongly agree* was recorded as a 1, so lower average SET scores imply more favorable evaluations. The responses to each question were highly correlated; the average Pearson correlation was 0.84. There was less variation in the SET dependent variable, partly because it was an average (mean = 1.52; std. dev. = 0.22). The mean SET score was the appropriate unit of analysis because teaching effectiveness at GVSU is evaluated by this measure. However, some universities have a specific question relating to teaching quality that is closely evaluated by administrators. To compare these results with other institutions, we estimated alternative specifications using individual questions.

The basic specification of our model was as follows:

$$(SET_{ijz} - \overline{SET_{ij}}) = F(ECG_{ijz} - \overline{ECG_{ij}}, X_{ijz} - \overline{X}_{ij}) + (\mu_{ijz} - \overline{\mu}_{ij}).$$

The subscript *i* represents a professor, *j* represents a course, and *z* represents a class. The dependent variable is average student teaching evaluation scores for a class (SET). *ECG* is the class' grade expectations. ***X*** is a vector of control variables that are included in every regression. Although our fixed-effects model controlled for instructor and class-specific effects, other class differences may still exist for the same instructor, teaching the same course over time. The following control variables are in the ***X*** vector:

1. Class size (Class Size)
2. Percentage of students taking a required course (Percent Core)
3. Percentage of students that are majors (Percent Major)
4. Average cumulative GPA of students in each class (Cumulative GPA)
5. For a class that has a special designation as a writing across the curriculum course, thereby including intensive writing requirements (Writing = 1)
6. Control variable = 1 for length of class 50 minutes, 75 minutes, or 150 minutes, respectively (Short, Medium, or Long)
7. Four control variables = 1 for classes that begin before 10 a.m., 10 a.m.–2 p.m., 2 p.m.–5 p.m., and after 5 p.m., respectively (Morning, Midday, Afternoon, or Evening)
8. Control variable = 1 for location of a class (Main Campus, Campus 2, or Campus 3)
9. Percentage of class that is represented in SET (Percent Responding)
10. Number of years a faculty member has taught at GVSU (Service Time)

To incorporate faculty experience and to control for time-varying factors, we employed as a control variable the number of years that a faculty member had been at the school at the time the class was taught. Anecdotal evidence suggests that faculty members adjust their teaching and grading during the first year of teaching. A Chow test indicated that the observations for the first year of teaching at GVSU are significantly different at the 1 percent level from the rest of the data set ($F = 3.69$). Consequently, we excluded classes taught in the instructor's first year from all our model specifications. An alternative specification employing dummy variables for each year yielded similar results.

We examined the distribution of the courses to ensure that data for the class-specific SET scores and expected grades were well distributed across the different control variables and that there were no significant outliers skewing the results.

TABLE 1. Descriptive Statistics: Economics and Finance Classes

Variable	Mean	Std. dev.	Variable	Mean	Std. dev.
SET	1.515	0.222	Afternoon	0.219	0.415
Expected Grade	3.156	0.248	Evening	0.269	0.444
Cumulative GPA	2.950	0.166	Main Campus	0.715	0.452
Relative Expected Grade	0.206	0.269	Campus 1	0.254	0.436
			Campus 2	0.031	0.173
Class Size	35.015	9.164	Pace of Class	3.212	0.425
150 minute class	0.327	0.470	Difficulty of Class	2.590	0.487
75 minute class	0.473	0.500	Percent Responding	0.767	0.130
50 minute class	0.200	0.401	Percent Core	0.320	0.158
Morning	0.119	0.325	Percent Major	0.166	0.209
Midday	0.392	0.489	Service Time	8.423	8.540

RESULTS

The summary statistics for each variable in the economics and finance departments are provided in Table 1. In the spirit of sensitivity analysis, we estimated and tested using a variety of specifications.

Difference of Means Test

We began the analysis with a simple difference in means test. To obtain observations that were comparable, we eliminated classes in satellite campuses and classes with special writing requirements. Initially, student grade expectations were proxied by expected grade (question 31 in the survey). For each instructor, we took the same course that an instructor had taught several times and picked the class with the highest expected grade and the class with the lowest expected grade in this course and then calculated the difference in the two SET scores. We performed a *t* test on the differential between two SET scores of the same courses taught by the same instructor. The mean difference was 0.10 with a *t* value of 2.53, which was significant at .01 (Type I error level).

Simple Model

We estimated a simple model that included expected grade, class size, and the average cumulative GPA of the class (model 1). Class size was significant in most studies (Danielsen and White 1976; Mirus 1973; Becker and Powers 2001). The cumulative average GPA of the students' controlled for the student draw that was represented in each class.

The results presented in Table 2, model 1, indicate that the coefficient for Expected Grade is significant at the 1 percent level (t value $= 4.17$). The negative sign indicates that as the average expected grade in a course goes up, the average SET improves (moves closer to one). The Class Size variable was also significant at the 1 percent level (t value $= 4.94$).

TABLE 2. Models Using Economics and Finance

	Model 1	Model 2	Model 3	Model 4
Expected Grade	−0.248	−0.207	−0.211	
	(4.17)**	(3.10)**	(3.25)**	
Cumulative GPA	0.235	0.190	0.192	
	(2.46)*	(1.86)$^{+}$	(1.91)$^{+}$	
Relative Expected Grade				−0.207
				(3.37)**
Class Size	0.007	0.004	0.004	0.004
	(4.94)**	(1.99)*	(1.98)*	(1.99)*
150 minute class		−0.122	−0.124	−0.124
		(2.25)*	(2.55)*	(2.55)*
75 minute class		−.00004		
		(0.00)		
Midday Class		−0.009		
		(0.24)		
Afternoon Class		0.002		
		(0.05)		
Evening Class		0.100	0.108	0.107
		(1.57)	(2.17)*	(2.16)*
Campus 1		−0.036	−0.037	−0.037
		(1.19)	(1.25)	(1.26)
Campus 2		−0.134	−0.134	−0.135
		(1.89)$^{+}$	(1.94)$^{+}$	(1.97)*
Pace of Class		−0.095	−0.092	−0.092
		(1.95)$^{+}$	(1.91)$^{+}$	(1.91)$^{+}$
Difficulty of Class		0.068	0.066	0.065
		(1.45)	(1.43)	(1.43)
Percent Responding		−0.140	−0.138	−0.138
		(1.60)	(1.60)	(1.61)
Percent Core		−0.047		
		(0.42)		
Percent Major		−0.146	−0.140	−0.141
		(1.21)	(1.20)	(1.22)
Service Time		0.010	0.010	0.010
		(1.19)	(1.33)	(1.32)
Constant	1.361	1.706	1.680	1.629
	(4.57)**	(5.04)**	(5.24)**	(9.62)**
Observations	260	260	260	260
Class/professor groups	44	44	44	44
Within R^2	0.16	0.22	0.22	0.22

Note: Absolute value of *t* statistics in parentheses ($^{+}$significant at .10, *significant at .05, **significant at .01 Type I error level). Within *R*-square measures the proportion of the variation within instructor/course explained by the regressions (Gould 1996).

The Cumulative GPA variable was positive and significant. Note that this average GPA was for all the classes taken *before* this class. This coefficient implies that classes with a better student draw tend to be tougher on their instructors and provide less favorable SET. One explanation could be that students who have achieved a higher cumulative GPA attribute their expected grade more to their own learning ability; consequently they are less likely to reward the instructor with a favorable SET. Gigliotti and Buchtel (1990), Greenwald (1980), and Feldman (1997) have summarized psychological attribution studies that discuss these types of results.

Another possible explanation, which we explore in the next section, is that in fact the SET is determined by the difference between the students' expected grades and the grades they have been accustomed to obtaining.

Benchmark Regression Results

Theory does not provide a firm guide as to whether the specific instructor- and course-related effects are fixed or random. As an alternative to controlling for fixed-teacher and course effects, one could rationalize that each instructor or course has the same average impact on SET, subject to an additional error term that differs for each individual instructor or course. In this case, a random-effects model may be more suitable. A Hausman test indicates that the coefficients of the fixed-effects model estimated by OLS are systematically different from a random effects at generalized least squares (GLS) estimator at the 5 percent level (Hausman 1978). Given that the OLS fixed-effects model is consistent, the test indicates that the random effects GLS estimator does not adequately model the results.

In model 2, we extended the specification to include all the other already discussed control variables. Again, in model 2, we controlled for both instructor- and class-specific effects and excluded data for the instructor's first year of teaching. We did not employ a quadratic term for expected grade in any specification because it was found to be statistically insignificant. In addition, a Cook-Weisberg test for heteroscedasticity (Cook and Weisberg 1983) did not reject the null hypothesis of constant variance at the 10 percent level (chi-square = 0.57).

Note that the coefficient for Expected Grade was significant at the 1 percent level (*t* value = 3.10). In model 2 we controlled for a variety of factors. The data indicated that if the same instructor taught the same course and generated expectations of a higher grade from C to B, the SET improved by 0.21. Consistent with previous literature, Class Size was a significant determinant of SET. A class smaller by 10 students improved the SET by 0.04.

Consider the issue about differences in the difficulty of each class. We asked two questions in the SET: (1) How difficult was this course? (2) Was the pace of this course too fast? These questions were not part of the average SET score. The Pace of Class variable was statistically significant (*t* = 1.95) at the 10 percent level. The negative sign indicated that a course taught at a faster pace (closer to one) worsened the SET scores (closer to five). The Difficulty of Class variable was statistically insignificant, perhaps because a professor was not likely to vary difficulty for different classes of the same course, and we controlled for the student draw by the cumulative incoming GPA. However, the Pace variable was marginally significant because the speed of the classes had more variation.

Nonresponse students were those who were enrolled in the course but did not respond to the SET. We constructed a composite variable for the proportion (relative to the class size at the beginning of the semester) of students who were not responding either because of withdrawal or because they were absent on the day of the evaluations (Percent Responding). The percentage of students that responded to the SET was statistically insignificant ($t = 1.60$). We could not directly incorporate students who did not respond to the SET in our analysis. These results indicated that the effect of attrition was not likely to change the relationship between GPA and SET scores. Also, a long class tended to improve the SET.

In model 3, we dropped the variables that did not have a t value of more than one in absolute value. The joint F test of the insignificant variables was $F(4, 200) = 0.08$; consequently, we preferred the more parsimonious model 3. The evening class then resulted in less favorable SET at the 10 percent level. The significance pattern of the other control variables remained the same. Note that in all models reported so far the coefficient on GPA was approximately the negative of the coefficient on expected grade. A test of the null hypothesis that these two coefficients were equal in magnitude but opposite in sign cannot be rejected [$F(1, 204) = 0.04$].

In model 4, we replaced Expected Grade and Cumulative GPA with the difference between the two. This new variable, henceforth referred to as the Relative Expected Grade, was significant at the 1 percent level. The negative sign means that as expected grade increased relative to cumulative GPA, an instructor received a more favorable SET.

Empirical studies traditionally include expected grade as an explanatory variable. However, our results indicated that Relative Expected Grade was the preferred explanatory variable. We employed a nonnested J test to examine this specification issue. The null hypothesis that expected grade alone was the correct specification was rejected (t value $= 2.08$), and the null hypothesis that the difference between expected grade and cumulative GPA was the correct specification was not rejected (t value $= 0.21$). Similar results were obtained using the Cox-Pesaran test (Greene 1993, 222–25).

We tested for simultaneity bias in model 4 by estimating a two-stage least squares model and performing a Hausman test. The two-stage least squares model with actual grade and cumulative GPA as instruments for the Relative Expected Grade variable provided similar results as the OLS model. The t statistics for the predicted relative grade was -2.24. The Hausman test assessed if the OLS estimates were significantly different from the instrument variable estimates. The null hypothesis indicating that the model was generated by an OLS process could not be rejected at the 10 percent level (chi-square $= 0.37$). Because we were unable to reject the hypothesis that the OLS and the instrument-variable model were different, we preferred the more efficient OLS estimator.

The size of the Relative Expected Grade gap could have an asymmetric relationship with SET scores. To explore if the relationship between Relative Expected Grade and SET was linear, we estimated a kernel regression for model 4 (Salgado-Ugarte, Shimizu, and Taniuchi 1996). In this nonparametric procedure, an initial regression with all the explanatory variables in model 4 (other than Relative Expected Grade) was estimated (Figure 1). Subsequently, the residuals (the variation in SET scores not attributed to the other control variables) were correlated with Relative Expected Grade. This allowed us to investigate the

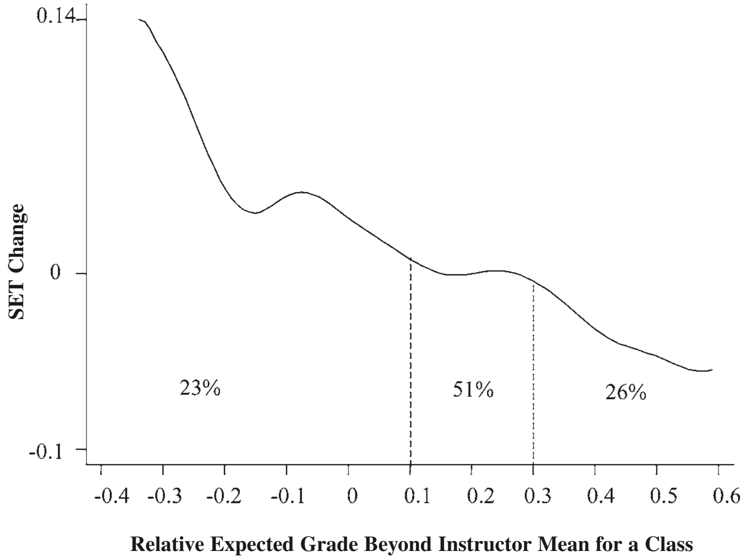

FIGURE 1. Kernel regression of SET and relative expected grade.

Note: Both variables are deviations from the mean and control for other explanatory variables. Zero on the Y axis represents average SET score of an instructor in a specific class. Zero on the X axis represents average Relative Expected Grade of an instructor in a specific class. Kernel regression shown uses a bandwidth of .06 and a Gaussian weight function. The results are not sensitive to reasonable bandwidth changes.

relationship without imposing a specific functional form on the data. The plot of the kernel regression estimates indicated that the relationship between Relative Expected Grade and SET scores was essentially linear with two moderate bumps. No particular low-order functional form was evident.

Note that both variables in the graph are deviations from the mean of the instructor for a specific course. The results indicated that initially there were large gains in SET scores when Relative Expected Grade gap approached the average. The most effective zone for Relative Expected Grade appeared to be when the curve flattened out at about 0.1 to 0.3 above the average. Approximately 51 percent of the courses taught by the faculty were in this zone, indicating that most faculty members have learned to be at the optimal range. Although there are some gains to be made when the Relative Expected Grade is higher than 0.3, the gains have to be balanced with the cost of being perceived as one who inflates grades.

Generalizing the Results

We considered extending the results to other schools by different specifications of the SET variable. Our dependent variable was mean SET scores for 25 questions. To compare our results to schools that focus on a single question related to teaching effectiveness, we estimated three alternative specifications. First, we estimated the model with the average of questions 20 to 25 that relate directly to teaching effectiveness (TE). Second, we estimated the model with two individual questions, question 20 (The instructor was able to make the material understandable) and question 22 (The instructor encouraged students to think for themselves). All three alternative specifications had a significant coefficient for Relative Expected Grade at the 1 percent level.

Generally, the variability of the SET scores increased as we moved from the overall average SET (std. dev. = 0.23) to an average SET for questions 20 to 25 (std. dev. = 0.27) and to individual questions such as no. 20 (std. dev. = 0.40). Our results indicated that the coefficients for Relative Expected Grade tended to get larger as we narrowed the dependent variable to more specific questions compared with the mean SET score for all the 25 questions. (Details of the results are available from the authors.) Compared with a Relative Expected Grade coefficient of −.21 for the average of 25 questions, the coefficient for the average of questions 20 to 25 was −.24, and the coefficients for questions 22 and 20 were −.27 and −.33, respectively.

The significance pattern across different model specifications indicated that the coefficient for Relative Expected Grade was robust. Our coefficients for Relative Expected Grade ranged in absolute value from 0.21 to a 0.33. If the incoming GPA was regarded as constant, these marginal estimates could be directly compared with the expected grade coefficients in the previous literature. Recall that Nichols and Soper (1972) obtained a value of 0.53 whereas Nelson and Lynch (1984) found it to be 0.15, both on a 4.0 scale. If these coefficients are standardized on a 5.0 scale, the Nichols and Soper value is 0.63 and the Nelson and Lynch value is 0.19. Consequently, our best estimates on a 5.0 scale (ranging between 0.21 to 0.33) are closer to the Nelson and Lynch estimate and considerably lower than the Nichols and Soper values.

CONCLUSIONS

It is important to note that the impact of Relative Expected Grade on SET is consequential in terms of faculty ranking. When the annual average ranking of the SET scores of the 58 faculty members are ordered, the SET scores are compressed around the median value (1.60), although the values can range between 1 and 5. On the basis of an average coefficient gap of 0.21, a change from an average grade of C+ to B would move the relative ranking of a faculty member with an average SET above 13 other faculty members (or above 21 percent of the business school faculty).

In this study, we used course-specific data that were confined to the observations from one school. Obviously our results may not translate directly to other schools. At GVSU, SET plays a major role in determining faculty compensation. Thus, our estimates of the coefficient for the Relative Expected Grade variable may generalize more easily to schools that effectively link SET scores to the determination of faculty salaries. Other schools that do not consider SET scores as an important criterion for faculty evaluation (research output may be the dominant criterion) or that have fixed salary increases that are not closely linked to annual performance (e.g., because of union contracts) may well have a different relationship.

Recall that the Seiver (1983) and Nelson and Lynch (1984) studies did not find a significant relationship between SET and expected grades for student-specific data in a simultaneous framework. Evidence of feedback from SET to expected grades is more likely when individual students are compared across instructors. Our alternative approach, based on class-specific data, did not indicate a simultaneous bias and showed a significant relationship. A simultaneous bias is less likely to exist between average expected grades and average SET in class-specific data. Moreover, the fixed-effects model controls for any time-invariant instructor or course effects.

The basic relationship between Relative Expected Grade and SET seems robust in our sample. Our results indicate that the expected grade relative to the incoming GPA of students provides more explanatory power. Inclusion of a wide variety of control variables, instructor- and course-related fixed effects, and difficulty and nonresponse bias variables, does not change the result that Relative Expected Grade is a significant determinant of SET. Further studies are needed to assess if these results are robust across different evaluation regimes.

APPENDIX
SET SURVEY INSTRUMENT

Instructor Evaluation

Please indicate your answers to the questions by filling in the appropriate circle on the answer sheet. Use the following scale A = *Strongly agree* and E = *Strongly disagree* for questions 1 through 28. If you have no opinion regarding any question, or if the question is not applicable to your class, please choose F.

ORGANIZATION/PREPARATION

1. A complete and logical syllabus was presented.
2. The presented syllabus was followed.
3. The course and lectures were well organized.
4. The instructor promptly returned homework, tests, and papers.
5. The instructor began and ended classes on time.
6. The instructor was well prepared for class.
7. The instructor explained the purpose of each class/lesson.
8. The instructor posted and kept office hours.

KNOWLEDGE

9. The instructor was able to supplement the material with relevant examples.
10. The instructor was able to answer student questions.
11. The instructor had current knowledge about the subject.

CLASSROOM ATMOSPHERE

12. I felt able to ask questions or express opinions.
13. Considering class size and content, the instructor was able to obtain an appropriate level of class participation.

14. The instructor was able to effectively lead and manage the class.
15. The instructor seemed genuinely concerned with student learning/progress.

FAIRNESS

16. The instructor explained to students how they would be evaluated/graded.
17. The instructor stuck to the grading plan presented.
18. Examinations reflected the material the instructor identified as important.
19. There were enough graded assignments (homework, tests, papers, projects) to reflect what was learned.

TEACHING EFFECTIVENESS

20. The instructor was able to make the material understandable.
21. The level at which the course was taught was appropriately challenging.
22. The instructor encouraged students to think for themselves.
23. The tests/papers/assignments encouraged thinking and reasoning.
24. The assignments and/or outside readings were useful.
25. The instructor gave helpful feedback on papers and exams.

COURSE CHARACTERISTICS

26. Relative to other courses, this course was very difficult.
27. Relative to other courses, this course's workload was very light.
28. Relative to other courses, this course's pace was too fast.

STUDENT CHARACTERISTICS

29. My major is (Accounting = A; Economics = B; Finance = C; Management = D; Marketing = E; General Business = F; Other = G).
30. My reason for taking this course is (Required course for major = A; Core course = B; Economics cognate requirement = C; Elective = D; Other = E).
31. My expected grade in the course is (A = A; B = B; C = C; D = D; F = F).

REFERENCES

Becker, W., and J. Powers. 2001. Student performance, attrition, and class size given missing student data. *Economics of Education Review* 20 (4): 377–88.

Becker, W., and M. Watts. 1999. How departments of economics evaluate teaching. *American Economic Review Proceedings* 89 (2): 344–49.

Cook, R. D., and S. Weisberg. 1983. Diagnostics for heteroscedasticity in regression. *Biometrika* 70 (1): 1–10.

Danielsen, L., and A. White. 1976. Some evidence on the variables associated with student evaluations of teaching. *Journal of Economic Education* 7 (Spring): 117–19.

DeCanio, S. 1986. Student evaluations of teaching: A multinomial logit approach. *Journal of Economic Education* 17 (3): 165–76.

Feldman, K. A. 1997. Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry and J. C. Smart, eds., *Effective teaching in higher education: Research and practice.* 368–95. New York: Agathon Press.

Gigliotti, R., and F. Buchtel. 1990. Attributional bias and course evaluations. *Journal of Educational Psychology* 82 (2): 341–51.

Gillmore, G., and G. Greenwald. 1999. Using statistical adjustment to reduce biases in student ratings. *American Psychologist* 54 (7): 518–19.

Gould, W. 1996. Why isn't the calculation of R^2 the same for areg and xtreg, fe? *STATA FAQ Statistics*, http://www.stata.com/support/faqs/stat/xtr2.html.

Greene, W. 1993. *Econometric analysis.* 222–25. Englewood Cliffs: Prentice-Hall.

Greenwald, A. 1980. The totalitarian ego: Fabrication and revision of personal history. *American Psychologist* 35 (7): 603–18.

———. 1997. Action editor, validity concerns and usefulness of student ratings of instruction. *American Psychologist* 52 (11): 1182–225.

Hausman, J. 1978. Specification tests in econometrics. *Econometrica* 46 (6): 1252–72.

Marsh, H. 1987. Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11 (3): 263–353.

Mirus, R. 1973. Some implications of student evaluation of teachers. *Journal of Economic Education* 5 (Fall): 35–37.

Nelson, J., and K. Lynch. 1984. Grade inflation, real income, simultaneity, and teaching evaluations. *Journal of Economic Education* 15 (1): 21–37.

Nichols, A., and J. Soper. 1972. Economic man in the classroom. *Journal of Political Economy* 80 (September/October): 1069–73.

Salgado-Ugarte, I., M. Shimizu, and T. Taniuchi. 1996. Nonparametric regression: Kernel, WARP, and k-NN estimators. *STATA Technical Bulletin* 30 (March): 15–31.

Seiver, D. 1983. Evaluations and grades: A simultaneous framework. *Journal of Economic Education* 14 (3): 33–38.

Watts, M., and W. Bosshardt. 1991. How instructors make a difference: Panel data estimates from principles of economics courses. *Review of Economics and Statistics* 73 (2): 336–40.

Call for Papers

The National Council on Economic Education and the National Association of Economic Educators will conduct three sessions at the January 2006 meetings of the Allied Social Science Association (ASSA) in Boston, MA. New research papers on any relevant topic in economic education will be considered. Those interested in presenting a paper should send an abstract, or the complete paper, no later than May 1, 2005, to Paul W. Grimes, Professor of Economics, Mail Stop 9580, Mississippi State University, Mississippi State, MS 39762–9580. Alternatively, papers and/or expressions of interest in serving as a discussant may be sent via email to pgrimes@cobilan.msstate.edu.

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Economics of Education Review 24 (2005) 369–376

Economics of Education Review

www.elsevier.com/locate/econedurev

Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity

Daniel S. Hamermesh*, Amy Parker

Department of Economics, University of Texas, Austin, TX 78712-1173, USA

Received 14 June 2004; accepted 21 July 2004

Abstract

Adjusted for many other determinants, beauty affects earnings; but does it lead directly to the differences in productivity that we believe generate earnings differences? We take a large sample of student instructional ratings for a group of university teachers and acquire six independent measures of their beauty, and a number of other descriptors of them and their classes. Instructors who are viewed as better looking receive higher instructional ratings, with the impact of a move from the 10th to the 90th percentile of beauty being substantial. This impact exists within university departments and even within particular courses, and is larger for male than for female instructors. Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible.
© 2004 Elsevier Ltd. All rights reserved.

JEL classifications: J71; I29

Keywords: Beauty; Discrimination; Class evaluations; College teaching

> It was God who made me so beautiful. If I weren't, then I'd be a teacher.
>
> [Supermodel Linda Evangelista]

1. Introduction

An immense literature in social psychology (summarized by Hatfield & Sprecher, 1986) has examined the impact of human beauty on a variety of non-economic outcomes. Recently economists have considered how beauty affects labor market outcomes, particularly earnings, and have attempted to infer the sources of its effects from the behavior of different economic agents (Hamermesh & Biddle, 1994; Biddle & Hamermesh, 1998). The impacts on these monetary outcomes are implicitly the end results of the effects of beauty on productivity; but there seems to be no direct evidence of the impacts of beauty on productivity in a context in which we can be fairly sure that productivity generates economic rewards.

A substantial amount of research has indicated that academic administrators pay attention to teaching quality in setting salaries (Becker & Watts, 1999). A number of studies (e.g., Katz, 1973; Siegfried & White, 1973; Kaun, 1984; Moore, Newman, & Turnbull, 1998) have demonstrated that teaching quality generates ceteris paribus increases in salary (but see DeLorme, Hill, & Wood, 1979). The question is what generates the measured productivity for which the economic rewards

*Corresponding author. Tel.: +1 512 475 8526; fax: +1 512 471 3510.
E-mail address: hamermes@eco.utexas.edu (D.S. Hamermesh).

doi:10.1016/j.econedurev.2004.07.013

are being offered. One possibility is simply that ascriptive characteristics, such as beauty, trigger positive responses by students and lead them to evaluate some teachers more favorably, so that their beauty earns them higher economic returns.

In this study we examine the productivity effects of beauty in the context of undergraduate education.[1] In particular, we consider the impact of instructors' looks on their instructional ratings in the courses that they teach. In Section 2 we describe a data set that we have created to analyze the impact of beauty on this indicator of instructors' productivity. In Section 3 we discuss and interpret the results of studying these impacts. Section 4 presents the implications of the analysis for interpreting the impact of an ascriptive characteristic on economic outcomes as stemming from productivity effects or discrimination.

2. Measuring teaching productivity and its determinants

The University of Texas at Austin, like most other institutions of higher learning in the United States and increasingly elsewhere too, requires its faculty to be evaluated by their students in every class. Evaluations are carried out at some point in the last 3 weeks of the 15-week semester. A student administers the evaluation instrument while the instructor is absent from the classroom. The rating forms include: "Overall, this instructor was very unsatisfactory (1); unsatisfactory (2); satisfactory (3); very good (4); excellent (5);" and "Overall, this course was very unsatisfactory, unsatisfactory …." In the analysis we concentrate on responses to the second question, both because it seems more germane to inferring the instructor's educational productivity, and because, in any event, the results for the two questions are very highly positively correlated ($r = 0.95$).

We chose instructors at all levels of the academic hierarchy, obtaining instructional staffs from a number of departments that had posted all faculty members' pictures on their departmental websites. An additional ten faculty members' pictures were obtained from miscellaneous departments around the University. The average evaluation score for each undergraduate class that the faculty member taught during the academic years 2000–2002 is included. This sample selection criterion resulted in 463 classes, with the number of classes taught by the sample members ranging from 1 to 13. The classes ranged in size from 8 to 581 students (enrolled as of the 12th day of the semester, after which it becomes costly to drop a class or even switch sections in a multi-section course), while the number of students completing the instructional ratings ranged from 5 to 380. Underlying the 463 sample observations are 16,957 completed evaluations from 25,547 registered students. Both lower- and upper-division courses are included. We make this distinction because there is no way of knowing the fraction of students in a particular course for whom it is required, which might otherwise be more interesting.

We also obtained information on each faculty member's sex, whether on the tenure track or not, minority status and whether he/she received an undergraduate education in an English-speaking country.[2] Table 1 presents the statistics describing these variables and the information about the classes. The means are weighted (by the number of evaluation forms returned) averages of the individual class averages. These descriptive statistics are generally unsurprising: (1) the average class rating is below that for the instructor him/herself; (2) the average rating is around 4.0 (on the 5 to 1 scale), with a standard deviation of about 0.5; and (3) non-tenure track faculty are disproportionately assigned to lower-division courses.

Each of the instructors' pictures was rated by each of six undergraduate students: three women and three men, with one of each gender being a lower division, two upper-division students (to accord with the distribution of classes across the two levels). The raters were told to use a 10 (highest) to 1 rating scale, to concentrate on the physiognomy of the instructor in the picture, to make their ratings independent of age, and to keep 5 in mind as an average. In the analyses we unit normalized each rating. To reduce measurement error the six normalized ratings were summed to create a composite standardized beauty rating for each instructor.

Table 2 presents statistics describing the ratings of the instructors' beauty by each of the six undergraduates who did the ratings. The students clearly had some difficulty holding to the instruction that they strive for an average rating of 5, as the averages of three of the six raw ratings were significantly below that, and none was significantly above (perhaps reflecting the students' inability to judge these older people, perhaps reflecting the choices implied in the epigraph). Moreover, the standardized ratings show that five of the six sets of

[1] Linking instructors' looks to their pedagogical productivity does not appear to have been done previously, but Goebel & Cashen (1979) and Buck & Tiene (1989) did ask students in various grades to comment on the teaching ability that they would expect from individuals of varying levels of beauty based on a set of photographs. Ambady & Rosenthal (1993), the only study to look at actual teaching evaluations (of 13 TAs in a single course), focused on their non-verbal behavior but did touch on the effects of their attractiveness.

[2] This last variable is designed to account for the possibility of lower productivity of foreign teachers (see Borjas, 2000, but also Fleisher, Hashimoto, & Weinberg, 2002) that might also be correlated with perceptions of their looks. In fact, in our sample this correlation is only −0.02.

Table 1
Descriptive statistics, courses, instructors and evaluations

Variable	All	Lower division	Upper division
Course evaluation	4.022 (0.525)	4.060 (0.563)	3.993 (0.493)
Instructor evaluation	4.217 (0.540)	4.243 (0.609)	4.196 (0.481)
Number of students	55.18 (75.07)	76.50 (109.29)	44.24 (45.54)
Percent evaluating	74.43	73.52	74.89
Female	0.359	0.300	0.405
Minority	0.099	0.110	0.090
Non-native English	0.037	0.007	0.060
Tenure track	0.851	0.828	0.869
Lower division	0.339	—	—
One credit	0.029	—	—
Number of courses	463	157	306
Number of faculty	94	42	79

Note: Means with standard deviations in parentheses. All statistics except for those describing the number of students, the percent evaluating the instructor and the lower–upper division distinction are weighted by the number of students completing the course evaluation forms.

Table 2
Beauty evaluations, individual and composite

	Average	Standard deviation	Standardized	
			Minimum	Maximum
Individual ratings:				
Male, upper division—1	4.43	2.18	−1.57	2.10
Male, upper division—2	4.87	1.65	−2.34	2.50
Female, upper division—1	5.18	2.05	−2.03	1.84
Female, upper division—2	5.39	2.10	−2.10	2.20
Male, lower division	3.53	1.70	−1.49	2.04
Female, lower division	4.14	1.88	−1.67	2.05
Composite standardized rating				
	0	0.83	−1.54	1.88

ratings were skewed to the right. There was some concern, based on observations in earlier research, that the distribution of ratings of female faculty might have higher variance than that of males. While the variance was slightly higher, the Kolmogorov–Smirnov statistic testing equality of the two distributions had a p-value of 0.077.

Despite these minor difficulties, a central concern—that the assessments of beauty be consistent across raters—was achieved remarkably well. The 15 pairwise correlation coefficients of the standardized beauty ratings range from 0.54 to 0.72, with an average correlation coefficient of 0.62. Cronbach's alpha, the standard psychometric measure of concordance, is 0.91. These indicate substantial agreement among the raters about the looks of the 94 faculty members. Any disagreement or greater subjectivity about the ratings would, however, merely impart a downward bias to estimates of the impact of beauty on teaching evaluations.

3. Impact of beauty on teaching ratings

3.1. Basic results

The basic model specifies a faculty member's teaching ratings as determined by a vector of his/her characteristics, X, and by a vector of the course's characteristics, Z. Included in X are whether the instructor is female, whether he/she is a minority, whether not a native English speaker, and whether on the tenure track. The central variable in X is our composite measure of standardized beauty. Z includes whether the observation is on an upper- or lower-division course, and whether it is for one credit. (27 of the classes were one-credit labs,

Table 3
Weighted least-squares estimates of the determinants of class ratings

Variable	All	Males	Females	Lower division	Upper division
Composite standardized beauty	0.275 (0.059)	0.384 (0.076)	0.128 (0.064)	0.359 (0.092)	0.166 (0.061)
Female	−0.239 (0.085)	—	—	−0.345 (0.133)	−0.093 (0.104)
Minority	−0.249 (0.112)	0.060 (0.101)	−0.260 (0.139)	−0.288 (0.156)	−0.231 (0.107)
Non-native English	−0.253 (0.134)	−0.427 (0.143)	−0.262 (0.151)	−0.374 (0.141)	−0.286 (0.131)
Tenure track	−0.136 (0.094)	−0.056 (0.089)	−0.041 (0.133)	−0.187 (0.141)	0.005 (0.119)
Lower division	−0.046 (0.111)	0.005 (0.129)	−0.228 (0.164)	—	—
One-credit course	0.687 (0.166)	0.768 (0.119)	0.517 (0.232)	0.792 (0.101)	—
R^2	.279	.359	.162	.510	.126
N courses	463	268	195	157	306
N faculty	94	54	40	42	79

Note: Robust standard errors in parentheses here and in Table 4.

physical education or other low-intensity activities that students tend to view differently from other classes).[3] Where sample sizes permit we examine the determinants of course evaluations in lower- and upper-division courses separately, since the students in the former may be more focused on the instructor him/herself and less on the degree to which the instructor can exposit the course material.

Table 3 presents weighted least-squares estimates of the equations describing the average course evaluations. As weights we use the number of students completing the evaluation forms in each class, because the error variances in the average teaching ratings are larger the fewer students completing the instructional evaluations. We present robust standard errors that account for the clustering of the observations (because we observe multiple classes for the overwhelming majority of instructors) for each of the parameter estimates.[4]

The striking fact from the estimates in the first column is the statistical significance of the composite standardized beauty measure. The effects of differences in beauty on the average course rating are not small: Moving from one standard deviation below the mean to one standard deviation above leads to an increase in the average class rating of 0.46, close to a one-standard deviation increase in the average class rating.[5] A complete picture of the importance of beauty in affecting instructors' class evaluations is presented in Fig. 1. For instructors at each percentile of the distribution of beauty, the figure shows the class evaluation that he/she would obtain with other characteristics in X and Z at the sample means. The instructional rating varies by nearly two standard deviations between the worst- and best-looking instructors in the sample.[6]

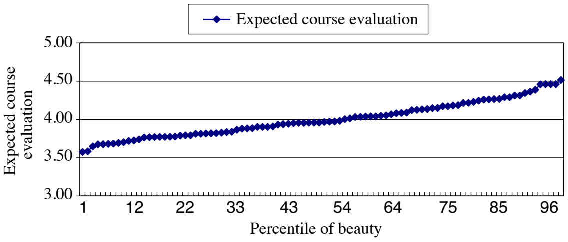

Fig. 1. Beauty and course evaluations.

[3] Age and a quadratic in age were included in other versions of the basic equation. These terms were never significantly nonzero as a pair or individually and had essentially no impact on the coefficients of the other terms in X and Z. Also unimportant was an indicator of whether the faculty member was tenured. If one-credit classes are excluded, the beauty coefficient changes slightly, rising to 0. 283.

[4] The unweighted least-squares parameter estimates differ little from those presented here. Had we failed, however, to use the correct robust standard errors, the parameter estimates here would all appear more highly significant statistically.

[5] This impact is at the intensive margin—among students who showed up in class on the day the course evaluations were completed. If we examine the extensive margin—the impact on the fraction of students attending class on that day—we also find a positive and nearly statistically significant effect of composite standardized beauty.

[6] One might be concerned about the upper and lower limits on the evaluation scores. While the lowest class average was 2.1, eight of the 463 classes did receive an average evaluation of 5.0. To examine whether this ceiling effect matters, we reestimated the basic equation in column 1 of Table 3 using an upper-limit tobit estimator. Not surprisingly, given the small fraction of observations at the ceiling, the parameter estimates were essentially unchanged. Least squares might also be problematic given the distribution of this measure. We thus also reestimated the basic equation using least absolute deviations. Again the coefficients were essentially unchanged (with the parameter estimate on the beauty measure rising slightly, to 0.299).

That inferring the impact of instructors' looks on measures of their instructional productivity requires evaluations of their looks by several raters is demonstrated by sequential reestimates of the basic equation that include each of the six raters' evaluations individually. While the class ratings are significantly related to each rater's views of the instructors, the estimated impacts range only from 0.12 to 0.23, i.e., below the estimates based on the composite standardized measure. There is substantial measurement error in the individual beauty ratings. The errors become less important once any pair of ratings is averaged: the estimated coefficients using the 15 possible pairs range from 0.19 to 0.26, and they range upward from 0.23 when any three ratings are averaged.

Minority faculty members receive lower teaching evaluations than do majority instructors, and non-native English speakers receive substantially lower ratings than do natives. Lower-division courses are rated slightly lower than upper-division courses. Non-tenure-track instructors receive course ratings that are surprisingly almost significantly higher than those of tenure-track faculty. This may arise because they are chiefly people who specialize in teaching rather than combining teaching and research, or perhaps from the incentives (in terms of reappointment and salary) that they face to please their students. The one-credit courses, all of which are lower-division, receive much higher evaluations than others, perhaps because of the nature of the courses as labs or electives.

Perhaps the most interesting result among the other variables in the vectors X and Z is the significantly lower rating received by female instructors, an effect that implies reductions in average class ratings of nearly one-half standard deviation. This disparity departs from the consensus in the literature that there is no relationship between instructor's gender and instructional ratings (Feldman, 1993).

To explore this sex difference further we estimate the basic model separately for classes taught by male and female instructors. The results are shown in columns 2 and 3 of Table 3. At the means of the variables the predicted instructional rating is lower for female instructors—the negative coefficient on the indicator in column 1 is not an artifact of a correlation of perceived beauty and gender. The reestimates show, however, that the impact of beauty on instructors' course ratings is much lower for female than for male faculty. Good looks generate more of a premium, bad looks more of a penalty for male instructors, just as was demonstrated (Hamermesh & Biddle, 1994) for the effects of beauty in wage determination.

Columns 4 and 5 show the results of estimating the equation separately for lower- and upper-division classes. The impact of beauty on instructional ratings, while statistically significant in both equations, is over twice as large in lower-division classes. Indeed, the same much bigger effects are found for two of the other variables that affected instructional ratings in the sample as a whole, whether the instructor is on the tenure track or is female. We might be tempted to conclude that class ratings by more mature students, and students who are learning beyond the introductory level in a subject, are less affected by factors such as beauty that are probably unrelated to the instructor's knowledge of the subject. Yet the impacts of being a minority faculty member or a non-native English speaker are just as large in the estimates for upper-division courses as in those for lower-division courses. It is unclear why the impacts of these variables among those in X are not attenuated in the more advanced courses. These estimates may imply the existence of discrimination by students in their evaluations, or they may result from shortfalls in the ability of those instructors to transmit knowledge or inspire students.

3.2. Robustness tests

One might be concerned that a host of statistical problems plagues the estimates shown in Table 3 and implies that our results are spurious. One difficulty is a potential measurement error: raters may be unable to distinguish physical attractiveness from good grooming and dress. Were this merely classical measurement error, we would have no difficulties. A subtle problem arises, however, if those who dress better, and whose photographs may thus be rated higher, are the same people who take care to be organized in class, to come to class on time, to hold their announced office hours, etc. What if our measure of beauty is merely a proxy for the general quality of the faculty member independent of his/her looks?

To account for this possibility we created an indicator equaling one for male faculty members who are wearing neckties in their pictures and for female faculty who are wearing a jacket and blouse. Formal pictures are on the websites of one-sixth of the faculty (weighted by numbers of students), and this indicator is added to a respecified version of the basic equation for which the results were shown in Column 1 of Table 3. The estimated impacts of this indicator and of composite standardized beauty are presented in the first row of Table 4. While instructors who present a formal picture do receive higher ratings, the inclusion of this additional measure reduces the estimated impact of beauty only slightly. The effect of composite standardized beauty remains quite large and highly significant statistically. We may conclude that the potential positive correlation of measurement error in the beauty ratings with unobservable determinants of teaching success does not generate serious biases in our estimates.

A related problem, also involving possibly non-classical measurement error, might arise if the more

Table 4
Alternative estimates of the relation between beauty and class ratings (lower- and upper-division classes)

		Variable				
		Composite standardized beauty	Formal dress	Black and white	Composite standardized beauty	
					Above mean	Below mean
1.	Photo bias (dress) ($N = 463$)	0.229 (0.047)	0.243 (0.088)			
2.	Photo bias (picture quality) ($N = 463$)	0.267 (0.063)		0.088 (0.106)		
3.	Photo bias (department) ($N = 414$)	0.236 (0.049)				
4.	Asymmetric beauty effect ($N = 463$)				0.237 (0.096)	−0.318 (0.133)
5.	Course fixed effects ($N = 157$)	0.177 (0.107)				

Note: The equations reported in rows 1–4 also include all the variables included in the basic equation in column 1 of Table 3. The equation reported in row 5 excludes variables in the vector Z.

concerned instructors were concerned enough about their pictures to include color rather than black-and-white photos on the websites. We classified the photographs along this criterion and again reestimated the basic equation. As the second row of Table 4 shows, there was almost no change in the parameter describing the relationship between composite standardized beauty and the evaluation. While the coefficient on the indicator variable "black-and-white" was small and statistically quite insignificant, it was somewhat surprisingly positive.[7]

Perhaps the most serious potential problem may result from a type of sample selectivity. Consider the following possibility. Among a group of people (a department), those who place their photographs on their websites will, until equilibrium in the game is reached, be better looking than those who do not present their photographs. They may also be people who are "go-getters" in other aspects of their lives, including their classroom teaching. If that is true, those instructors who are among the few in a department whose pictures are available will be better looking and be better instructors, while those from departments with all pictures available will on average be average looking and average instructors.

To examine this potential problem we reestimate the basic equation on the subsample of 84 faculty members, teaching 414 classes, in which an entire department's faculty's pictures are available. The results of estimating the basic equation over this slightly reduced sample are shown in the second row of Table 4. Compared to the basic estimate (0.275), accounting for this potential problem reduces the estimated impact of composite standardized beauty slightly and implies that a two-standard deviation change in beauty raises the course rating by 0.39 (three-fourths of a standard deviation in course ratings). Apparently this kind of selectivity matters a bit, but it does not vitiate the basic result.

The next possibility does not represent a potential bias in the basic results, but rather asks whether they are masking some additional sample information. There is some indication (Hamermesh & Biddle, 1994; Hamermesh, Meng, & Zhang, 2002) that the effect of beauty on earnings is asymmetric, with greater effects of bad than of good looks. Does this asymmetry carry over into its effects on productivity in college teaching? To examine this possibility we decompose the composite standardized beauty measure into positive and negative values and reestimate the basic equation allowing for asymmetry. The results are shown in the third row of Table 4. The effect on course ratings of looking better than average is slightly below and opposite in sign of the effect of looking worse than average.[8] There is only slight evidence of asymmetry in the impact of instructors' beauty on their course ratings.

Another potential issue is that courses may attract students with different attitudes toward beauty. These may be correlated with the instructional ratings that the students give and may also induce departmental administrators to assign courses to instructors based on their looks. Some courses may also generate different ratings

[7] Yet another potential difficulty is that the photographs may not all be equally current. Given that all had to be in electronic files, and given the strong evidence (Hatfield & Sprecher, 1986, pp. 282–3) that an individual's perceived beauty changes very slowly with age, even a correlation between the age of the photograph and an instructor's evaluation would cause at most a minimal bias in any estimates.

[8] The t-statistic on the hypothesis that they are equal and opposite in sign is 0.41. This may not contradict results indicating asymmetric effects of beauty on earnings. Many more individuals are rated above average in looks than are considered below average, so that the asymmetry might not exist if the beauty measure itself were symmetric, as it is by construction here.

depending on their difficulty, their level and other differences, and these may be correlated with the instructor's looks. The gender mix of students may differ among courses, and this too may affect the estimated impacts of beauty. To examine these possibilities we take advantage of the fact that 157 of the 463 classes in our sample are instructed by more than one faculty member over the 2 years of observation. These courses involve 54 different instructors (of the 94 in the sample). We reestimate the basic equation on this subsample adding course fixed effects. Thus any estimated effect of beauty will reflect within-course differences in the impact of looks on instructional ratings.

The results are presented in the final row of Table 4. The estimated impact of composite standardized beauty on class evaluations is somewhat smaller than in the other estimates, but still substantial. This is mostly due to sampling variability. Reestimating the basic equation of Table 3 over this reduced sample of 157 classes yields an impact of composite standardized beauty on instructional ratings of 0.190 (s.e. = 0.079).[9]

4. Conclusion and interpretations

The estimates leave little doubt that measures of perceived beauty have a substantial independent positive impact on instructional ratings by undergraduate students. We have accounted for a variety of possibly related correlates, and we have shown that the estimated impacts are robust to potential problems of selectivity, correlated measurement error and other difficulties. The question is whether these findings really mean that beauty itself makes instructors more productive in the classroom, or whether students are merely reacting to an irrelevant characteristic that differs among instructors.

The first issue is that our measure of beauty may merely be a proxy for a variety of related unmeasured characteristics that might positively affect instructional ratings. To the extent that these are positively correlated with beauty but not caused by it, our results overstate the impact of beauty. That we have held constant for as many course and instructor characteristics as we have should mitigate some concerns about this potential problem. If there is a characteristic that is caused by a person's physical appearance and that also generates higher instructional ratings, then failing to measure it (and excluding it from the regressions) is correct. For example, if good-looking instructors are more self-confident because their beauty previously generated better treatment by other people, and if their self-confidence makes them more appealing instructors, it is their beauty that is the ultimate determinant of (part of) their teaching success.

A second and more important issue is whether higher instructional ratings mean that the faculty member is a better teacher—is more productive in stimulating students' learning. The instructional ratings may putatively reflect productivity, but do they really do so? Discussions of this question among administrators and faculty members have proceeded since instructional evaluation was introduced, and we do not wish to add to the noise. Regardless of the evidence and of beliefs about this issue, however, instructional ratings are part of what universities use in their evaluations of faculty performance—in setting salaries, in determining promotion, and in awarding special recognition, such as teaching awards. Thus even if instructional ratings have little or nothing to do with actual teaching productivity, university administrators behave as if they believe that they do, and they link economic rewards to them. Thus the ratings are at least one of the proximately affected outcomes of beauty that in turn feed into labor-market outcomes.

The most important issue is what our results tell us about whether students are discriminating against ugly instructors or whether they really do learn less (assuming that instructional ratings reflect learning). For example, what if students simply pay more attention to good-looking instructors and learn more from them? We would argue that this is a productivity effect—we would claim that the instructors are better teachers. Others might (we think incorrectly) claim that the higher productivity arises from students' (society's) treating them differently from their worse-looking colleagues and is evidence of discrimination. Disentangling the effects of differential outcomes resulting from productivity differences and those resulting from discrimination is extremely difficult in all cases, as we believe this unusual illustration of the impact of beauty on a physical measure that is related to earnings illustrates.

The epigraph to this study may be correct—someone who does not qualify to be a supermodel might well go into teaching. Even in college teaching, however, our evidence demonstrates that a measure that is viewed as reflecting teaching productivity, whether it really does so or not, is also one that is enhanced by the instructor's pulchritude.

Acknowledgements

We thank William Becker, Jeff Biddle, Lawrence Kahn, Preston McAfee, Alex Minicozzi, Gerald Oettinger, a referee and participants at seminars at several universities for helpful suggestions.

[9] If we include a vector of indicators for departments in the basic equation in Table 3, we find a somewhat larger effect than here, although one that is still smaller than that in the basic equation.

References

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*, 431–441.

Becker, W., Watts, M. (1999). How departments of economics evaluate teaching. *Papers and proceedings 89*, American Economic Association (pp. 355–359).

Biddle, J., & Hamermesh, D. (1998). Beauty, productivity and discrimination: lawyers' looks and lucre. *Journal of Labor Economics*, *16*, 172–201.

Borjas, G. (2000). Foreign-born teaching assistants and the academic performance of undergraduates. *Papers and proceedings 90*, American Economic Association (pp. 344–349).

Buck, S., & Tiene, D. (1989). The impact of physical attractiveness, gender, and teaching philosophy on teacher evaluations. *Journal of Educational Research*, *82*, 172–177.

DeLorme, C., Hill, R. C., & Wood, N. (1979). Analysis of a quantitative method of determining faculty salaries. *Journal of Economic Education*, *11*, 20–25.

Feldman, K. (1993). College students' views of male and female college teachers: part II. Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, *34*, 151–211.

Fleisher, B., Hashimoto, M., & Weinberg, B. (2002). Foreign GTAs can be effective teachers of economics. *Journal of Economic Education*, *33*, 299–326.

Goebel, B., & Cashen, V. (1979). Age, sex and attractiveness as factors in student ratings of teachers: a developmental study. *Journal of Educational Psychology*, *71*, 646–653.

Hamermesh, D., & Biddle, J. (1994). Beauty and the labor market. *American Economic Review*, *84*, 1174–1194.

Hamermesh, D., Meng, X., & Zhang, J. (2002). Dress for success: does primping pay? *Labour Economics*, *9*, 361–373.

Hatfield, E., & Sprecher, S. (1986). *Mirror, Mirror* .... Albany, NY: State University of New York Press.

Katz, D. (1973). Faculty salaries, promotions, and productivity at a large university. *Papers and proceedings 63*, American Economic Association (pp. 469–477).

Kaun, D. (1984). Faculty advancement in a nontraditional university environment. *Industrial and Labor Relations Review*, *37*, 592–606.

Moore, W. J., Newman, R., & Turnbull, G. (1998). Do academic salaries decline with seniority? *Journal of Labor Economics*, *16*, 352–366.

Siegfried, J., White, K. (1973). Financial rewards to research and teaching: a case study of academic economists. *Papers and proceedings 63*, American Economic Association (pp. 309–315).

The Journal of General Psychology, 2006, *133*(1), 19–35

Hot or Not: Do Professors Perceived as Physically Attractive Receive Higher Student Evaluations?

TODD C. RINIOLO
KATHERINE C. JOHNSON
TRACY R. SHERMAN
JULIE A. MISSO
Department of Psychology
Medaille College

ABSTRACT. Previous research investigating the influence of perceived physical attractiveness on student evaluations of college professors has been limited to a handful of studies. In this study, the authors used naturally occurring data obtained from the publicly available Web site www.ratemyprofessors.com. The data suggested that professors perceived as attractive received higher student evaluations when compared with those of a nonattractive control group (matched for department and gender). Results were consistent across 4 separate universities. Professors perceived as attractive received student evaluations about 0.8 of a point higher on a 5-point scale. Exploratory analyses indicated benefits of perceived attractiveness for both male and female professors. Although this study has all the limitations of naturalistic research, it adds a study with ecological validity to the limited literature.

Key Words: naturalistic research, physical attractiveness, student evaluations, teacher characteristics

OUR PURPOSE IN THE PRESENT STUDY was to investigate whether college professors perceived as physically attractive received higher student evaluations compared with colleagues that were perceived as nonattractive. To begin investigating this topic, we reviewed some relevant literature.

First, research results indicate that a variety of factors influence student evaluations of college professors (for pertinent reviews, see Cashin & Downey, 1992; Greenwald & Gillmore, 1997; Marsh & Roche, 1997; McKeachie, 1997). For

The authors thank Janice Laben and Raymond Bailey. They also acknowledge the helpful comments by the action editor, F. Richard Ferraro, and three anonymous reviewers.

Address correspondence to Todd C. Riniolo, Department of Psychology, Medaille College, 18 Agassiz Circle, Buffalo, NY 14214; triniolo@medaille.edu (e-mail).

example, in the Dr. Fox studies (Naftulin, Ware, & Donnelly, 1973; Ware & Williams, 1975), a professional actor (whose character was called Dr. Fox) was videotaped using high and low levels of expressiveness. When students viewed the taped lectures, Dr. Fox received higher evaluations when using the expressive style. Likewise, Williams and Ceci (1997), in a naturalistic study using an experienced professor who taught identical courses in the fall and spring semesters (a large section of Developmental Psychology), found an enthusiastic teaching style (while presenting the same course content) resulted in much higher student evaluations. Radmacher and Martin (2001), using hierarchical regression with a wide range of variables (professor's age and extraversion traits; student's current grades, gender, enrollment status, ACT scores, and age), found that professors' extraversion was the strongest predictor of midterm student evaluations of teaching effectiveness. However, students' enrollment status, current course grade, and age were also positively correlated with midterm student evaluations.

In addition, Freeman (1994), using three written descriptions of hypothetical professors (feminine, masculine, androgynous), found that students preferred both male and female professors who possess androgynous characteristics. Also, positive personality characteristics (e.g., caring, enthusiasm, sense of humor) were associated with higher student evaluations when undergraduates were asked to describe their best college professor (Basow, 2000; Waters, Kemp, & Pucci, 1988). Research results also indicate that perceived learning, prior interest in the subject (Marsh & Roche, 2000), students' expectations of grades (Greenwald & Gillmore, 1997; Millea & Grimes, 2002), nonverbal behavior (Ambady & Rosenthal, 1993), course workload (Marsh & Roche, 1997), and student motivation (Cashin & Downey, 1992) also influence student evaluations. In summary, the current evidence suggests that an array of factors, not just the quality of the course, impact student evaluations of their college professor.

Second, research results indicate that being perceived as physically attractive is associated with a wide range of positive outcomes (Bloch & Richins, 1992; Hosoda, Stone-Romero, & Coats, 2003; Langlois et al., 2000). For example, individuals perceived as attractive are more likely to receive help from strangers than are persons perceived as unattractive (Benson, Karabenick, & Lerner, 1976). In both simulated (Mazzella & Feingold, 1994) and real judicial trials (Stewart, 1980), defendants perceived as attractive were more likely to receive a more lenient punishment if found guilty of a crime. Researchers have also demonstrated that persons perceived as attractive (a) were viewed as more socially competent (Eagly, Ashmore, Makhijani, & Longo, 1991), (b) were viewed as having greater academic potential by teachers (Ritts, Patterson, & Tubbs, 1992), (c) were more persuasive communicators (Chaiken, 1979), and (d) were preferred by voters in political elections (Budesheim & DePaola, 1994; Sigelman, Thomas, Sigelman, & Ribich, 1986). Researchers have shown that individuals perceived as attractive receive higher incomes than co-workers perceived as unattractive (Frieze, Olson, & Russell, 1991; Hamermesh & Biddle, 1994). Hosoda et al.'s (2003) meta-analysis demonstrated that individuals perceived as attractive obtain better outcomes for a variety of job-related issues (e.g., hiring, promotion, performance evaluation). Although some factors, such as concern for others and integrity, have not demonstrated an influence of perceived attractiveness (Eagly et al.), the overall literature has indicated a wide variety of positive outcomes. In fact, Myers (2005) summarized this literature as, "Good looks are a great asset" (p. 432).

It is important to note that an individual's physical attractiveness is not an objective variable like heart rate or weight that can be measured with precise accuracy. Although there is general agreement about who is attractive both within and between cultures (Langlois et al., 2000), evaluating physical attractiveness is partially a subjective judgment (Eagly et al., 1991; Monin, 2003). Thus, individual raters can perceive and evaluate physical attractiveness somewhat differently. Furthermore, "there seems to be no agreed-upon criteria for defining physical attractiveness in attractiveness research" (Hosoda et al., 2003, p. 457). For the present study, the classification of "attractive" and "nonattractive" groups (i.e., professors) based on the perceptions of the majority of raters (i.e., students).

Moreover, rating attractiveness is not solely influenced by the physical appearance of the target (e.g., the professor) and individual preferences of the perceiver (e.g., the student), but additional influences can contribute. For example, the target's personality characteristics (Gross & Crofton, 1977), similarity of attitudes between perceiver and target (Klentz, Beaman, Mapelli, & Ullrich, 1987), the perceived familiarity of the target (Monin, 2003), the perceiver's sense of self (Horton, 2003), and the dating status and commitment to a partner in close relationships of the perceiver (D. J. Johnson & Rusbult, 1989; Simpson, Gangestad, & Lerma, 1990) all influence perceptions of attractiveness. Furthermore, in the majority of the physical attractiveness literature, researchers have relied upon first impressions. However, differences may exist between initial impressions compared with those following repeated exposures when the perceiver has additional information about the target (Eagly et al., 1991; Hosoda et al., 2003).

Research also indicates that other factors, such as the gender of the perceiver and the clothing of the target, also influence the perception of physical attractiveness (Abbey, Cozzarelli, McLaughlin, & Harnish, 1987; Buckley, 1983; Workman & Orr, 1996). For example, Williamson and Hewitt (1986) found that males perceived female models as more attractive in sexually alluring clothing, whereas women rated the female models as more attractive in neutral attire. Likewise, in studies investigating sexual harassment (K. P. Johnson & Workman, 1992) and acquaintance rape (Cassidy & Hurrell, 1995; Workman & Freeburg, 1999; Workman & Orr), researchers have also shown that clothing (e.g., skirt length) and gender of the perceiver can influence the perceptions of a target (i.e., the victim). Also, marketing researchers have demonstrated that adornments (e.g., makeup, hairstyle, jewelry) can also alter perceptions of physical attractiveness (Bloch & Richins, 1992; Mack & Rainey, 1990). In summary, the evaluation of who is perceived as physically attractive is not simply an objective variable, but is partially a subjective judgment that can be influenced by multiple inputs.

Currently, we are aware of only four studies (Ambady & Rosenthal, 1993; Buck & Tiene, 1989; Goebel & Cashen, 1979; Hammermesh & Parker, in press) in which researchers have attempted to investigate the influence of perceived physical attractiveness on student evaluations of college professors. In an initial study, Goebel and Cashen used 10 college freshmen to classify black-and-white photographs as attractive or unattractive. Twenty different freshmen subsequently judged presumed teaching effectiveness from the photographs. Results showed the photos judged as attractive received higher ratings. Buck and Tiene modified Goebel and Cashen's study with 42 undergraduate seniors by attaching a written statement about teaching philosophy (authoritarian or humanistic) to photos judged as attractive or unattractive. Results indicated no main effects of attractiveness on perceived competence, but interaction effects indicated that attractive female authoritarian photos received higher ratings compared with those of male (both attractive and unattractive) and unattractive female authoritarian photos. However, results of both studies have limited generalizability because the authors relied upon presumed (as opposed to real) student evaluations.

In the first study investigating the influence of perceived physical attractiveness using actual student evaluations, Ambady and Rosenthal (1993) primarily focused on the influence of nonverbal behavior. The authors asked two female undergraduates to rate attractiveness (5-point scale) from a still video clip of 13 graduate teaching fellows (6 women) who were teaching sections for undergraduate courses. The authors subsequently correlated perceived attractiveness ratings with real end-of-semester evaluations (comprised of the mean ratings from the students in the section). Perceived attractiveness was not statistically related to student evaluations ($r = .32$, *ns*), perhaps because of the low statistical power associated with the small sample size (n = 13). The results of that study are not only limited because of the small sample size of the teaching fellows, but have limited generalizability because perceived attractiveness was judged by only two female raters.

The most comprehensive investigation of this topic was recently performed by Hamermesh and Parker (in press). Six undergraduate students (3 women) rated the perceived attractiveness of 94 professors by using photographs posted on departmental Web sites. Physical attractiveness ratings (10-point Likert scale) were compared with the professors' real end-of-semester evaluations (number of students who completed student evaluations ranged from 5 to 380). Regression analysis indicated a strong influence of perceived attractiveness on student evaluations. Professors rated as attractive were more likely to receive higher evaluations. Subsequent analysis indicated that the influence of perceived attractiveness was stronger for male as compared with female professors. However, the results of that study were limited by the small number of students who rated perceived attractiveness.

In summary, the aforementioned literature on perceived physical attractiveness and student evaluations of college professors not only is restricted to a handful of studies, but is limited with respect to ecological (i.e., real-world) validity in several important ways. First, previous researchers did not use both the rankings of professors' perceived attractiveness and evaluations by students who were enrolled in the course. In addition, researchers who used real end-of-semester evaluations relied on very small numbers of students to rate attractiveness. Relying on ratings from a handful of undergraduates who were not enrolled in the course is potentially problematic and may have limited generalizability because evaluating attractiveness is partially a subjective judgment with multiple inputs (Eagly et al., 1991; Monin, 2003). Furthermore, previous researchers have relied upon perceptions of still images, which may differ from perceptions obtained by face-to-face interaction (Eagly et al.). As Buck and Tiene (1989) have noted, relying on a still image measures an initial impression. However, student evaluations are typically given at the end of the college semester after students have been repeatedly exposed to the professor. Thus, the perceiver may have different inputs for rating attractiveness between initial impressions and repeated exposures over time (Eagly et al.), a limitation that extends to most research on attractiveness (Hosoda et al., 2003).

Our purpose in the present study was to add to the limited literature by examining perceived physical attractiveness and student evaluations in a naturally occurring database of concurrent ratings. The universities shown in Table 1 have large numbers of student evaluations (as of June 1, 2004, ranging from 20,131 to 36,312) on the Internet Web site www.ratemyprofessors.com. A public Web site designed for students, www.ratemyprofessors.com posts anonymous and voluntary evaluations of college professors (the Web site is not university sponsored). Although our study has all the limitations of any research using naturalistic data (e.g., a potentially biased sample, lack of experimental control, the potential of multiple ratings), it adds to the literature a study with ecological validity. In this study, we compared professors rated as attractive with a nonattractive control group matched for department and gender. Furthermore, we performed multiple replications to determine if the results were statistically reliable (Riniolo & Schmidt, 2000). On the basis of the literature demonstrating that attractiveness is associated with many positive outcomes, we predicted that professors perceived as physically attractive would receive higher evaluations compared with colleagues perceived as nonattractive.

Method

Participants

In this study, we used student evaluations of professors from the Web site www.ratemyprofessors.com. We obtained evaluations on June 1, 2004, using the four schools with the most ratings (see Table 1). We selected the most rated

TABLE 1. Descriptive Statistics From www.ratemyprofessors.com

	Grand Valley State University[a]				University of Delaware[b]			
Category	*n*	%	*M*	*SD*	*n*	%	*M*	*SD*
All professors	1,714				2,000			
Total ratings	6,312				27,756			
All professors[e]	522				331			
Total ratings	25,590				15,599			
Student evaluations			3.56	0.82			3.50	0.86
Attractive[e]	80				45			
% attractive professors		15				14		
Student evaluations			4.22	0.46			4.11	0.74
No. of ratings			49.6	20.8			39.2	14.7
Hotness total/ no. of ratings			.25	0.17			.25	0.16
Nonattractive[e]	442				286			
Student evaluations			3.44	0.82			3.41	0.84
No of ratings			48.9	23.8			48.4	27.8

[a]Allendale, MI. [b]Newark, DE. [c]San Diego, CA. [d]Harrisonburg, VA. [e]≥ 25 ratings.

schools because (a) large numbers produced more precise estimates and have greater statistical power than smaller samples do (Cohen, 1988), (b) the large number of evaluations indicates that the Web site is well known and actively being used by students, and (c) replication is the best method for determining whether results are statistically reliable (Riniolo & Schmidt, 2000). Also, because just a single student rating will include a professor in the database, we limited data for subsequent statistical analysis to professors that had received at least 25 student evaluations. Researchers have demonstrated high reliability between class-average responses with at least 25 ratings (Marsh & Roche, 1997).

Descriptive information about the full pool of professors is provided in Table 1. Subsequent statistical comparisons between attractive and nonattractive groups (see description of matched analysis in a later section) included 156 professors (50 women, 32%) from Grand Valley State University, 90 professors (48 women, 53%) from the University of Delaware, 106 professors (32 women, 30%) from San Diego State University, and 48 professors (14 women, 29%) from James Madison University. Table 2 shows the number of departments represented.

San Diego State University[c]				James Madison University[d]			
n	%	*M*	*SD*	*n*	%	*M*	*SD*
2,205				1,214			
26,921				20,131			
285				275			
12,175				11,621			
		3.48	0.91			3.5	0.89
56				30			
	20				11		
		4.14	0.56			4.39	0.42
		47.5	28.1			38.0	13.9
		.34	0.23			0.27	0.15
229				245			
		3.32	0.90			3.41	0.88
		41.5	16.9			42.6	16.3

Given the naturalistic context of this study, limitations are evident, especially with regard to the ratings from which we took our data. In particular, the ratings made at the Web site are anonymous. Table 3 shows relevant information about undergraduate student populations of the four universities used in this study (i.e., the assumed populations). All have similar numbers of male and female students (range = 58%–60% female), but differences exist, such as percentage of minority and out-of-state students and scores on standardized tests (i.e., ACT and SAT).

Measures

On the Web site, raters calculate a professor's overall quality on a 5-point Likert-type scale (5 indicating the highest rating) by averaging how the instructor scores on helpfulness and clarity. Definitions of helpfulness and clarity are provided if raters "click" to receive further information. Helpfulness is defined on the Web site as: "This category rates the professor's helpfulness and approachability. Is the professor approachable and nice? Is the professor rude, arrogant, or just plain mean? Is the professor willing to help you after class?" Clarity is defined as: "How

TABLE 2. Departments Represented on www.ratemyprofessors.com

	Grand Valley State University		University of Delaware		San Diego State University		James Madison University	
Department	Men[a]	Women[b]	Men[a]	Women[b]	Men[a]	Women[b]	Men[a]	Women[b]
Accounting	1	1	1	0	1	0	N/A	N/A
Anthropology	1	0					1[a]	0[b]
Biology					0	1		
Business	2	2	1	1	2	1	1	2
Chemistry	1	0					1	0
Communication	1	0			2	1	1	1
Computer Science	2	0						
Criminal Justice	3	1	0	1				
Economics	1	0	3	2	5	0	1	0
Education	1	2	1	1	0	1		
English	6	3	5	5	7	4	2	0

Fine Arts	2	2					0	1
Geography					0	2		
Geology	0	1						
History	6	0	4	1	4	0	1	0
Hospitality	1	0						
Languages	2	3	1	8	2	1		
Math	3	2	1	1	4	1	0	1
Music	2	0						
Philosophy	3	3	0	1	4	0	4	0
Political Science	3	0	2	2	3	0	1	0
Psychology	6	3	1	0	2	2		
Science	5	0	1	0	1	0	1	0
Social Science	0	1	0	1				
Sociology	1	1			0	1	1	1
Theology							1	0
Writing							1	1

[a]Number of attractive/nonattractive male matches (each match represents two professors). [b]Number of attractive/nonattractive female matches (each match represents two professors).

TABLE 3. Undergraduate Information of Student Populations

Characteristic	Grand Valley State University[a]	University of Delaware[b]	San Diego State University[c]	James Madison University[d]
Students (*n*)	17,807	17,200	27,345	14,991
Full-time (%)	84	86	79	96
Women (%)	60	58	58	60
Applicants admitted (%)	73	42	50	62
Out of state (%)	4	58	7	30
Minority (%)	10.6	12.3	40.7	10.2
Largest minority	African American	African American	Hispanic American	Asian American or Pacific Islander
Live on campus (%)	29	50	48	40
Full-time freshman retention for 2002 (%)	78	90	82	92
ACT scores >24 (%)	46	71	39	NA
SAT verbal >600 (%)	NA	42	18	35
SAT Math >600 (%)	NA	54	26	39
Full-time faculty (%)	68	81	60	72
Student/faculty ratio	17:1	13:1	19:1	17:1

Source. *Four-year colleges*, 35th ed, by Peterson's, Princeton, NJ, 2004.
[a]Allendale, MI. [b]Newark, DE. [c]San Diego, CA. [d]Harrisonburg, VA.

well does the professor convey the class topics? Is the professor clear in his presentation? Is the professor organized and does the professor use class time effectively?" Perceived attractiveness (note that photos of professors do not appear on the Web site) is calculated from an optional appearance question. The Web site notes that this question is "just for fun," asking students whether their professor is "hot" or "not." The Web site calculates a hot, or not hot, rating (defined as a *hotness total*). Those with an equal or negative balance are assigned a zero rating (i.e., nonattractive), whereas positive balances are displayed (i.e., attractive). For group comparisons, those professors with a positive hotness total were classified as attractive (i.e., a majority who answered this optional question on the Web site perceived the professor as physically attractive). This attractiveness marker resulted in about 15% of the professors (with at least 25 ratings) being classified as attractive (see Table 1 for percentages at each university). Unfortunately, the Web site does not provide how many total hot or not ratings were cast. Table 1 shows the average hotness total divided by the total number of overall ratings for the professors perceived as attractive.

Procedure

We sorted data by number of ratings and included professors with at least 25 ratings in the sample. We divided the sample into attractive and nonattractive groups on the basis of the student ratings. We subsequently matched the sample for gender and department on the basis of attractive and nonattractive controls. In instances in which more than one potential control existed (i.e., a nonattractive professor of the same gender and department), we randomly selected the control from all potential matches. We only included instances of matched professors in the sample used for subsequent matched data analyses. To control for inflation of error rates, we limited the a priori planned analysis to the matched analysis just described.

Results

Descriptive statistics for all attractive and nonattractive professors, prior to matching, are provided in Table 1. After matching, independent *t* tests revealed statistically significant differences between groups because attractive professors had consistently higher evaluations compared with nonattractive controls (see Table 4). We subsequently converted the *t* values into Cohen's *d*, a measure of effect size in which 0.2 indicates a small difference between groups, 0.5 indicates a medium difference, and 0.8 indicates a large difference (Cohen, 1988). Results indicated a large effect size difference between groups (see Table 4).

We performed several exploratory analyses. First, we conducted separate analyses for male and female professors using the same participants in Table 4. Results showed that both attractive men and attractive women scored higher evaluations

TABLE 4. Student Evaluations After Controlling for Department and Gender

	Attractive		Nonattractive control				
School	*M*	*SD*	*M*	*SD*	*df*	*t*	Cohen's *d*
Grand Valley State	4.22	.46	3.39	.81	154	7.83**	1.25
University of Delaware	4.11	.74	3.44	.86	88	3.94**	0.83
San Diego State	4.13	.57	3.32	.82	104	5.93**	1.15
James Madison	4.41	.43	3.36	.86	46	5.35**	1.54

$^{**}p < .001$.

when compared directly against same-gender nonattractive controls (see Table 5). Second, we identified no statistical differences when comparing attractive men with attractive women at the same university (as shown in Table 2, departments varied between genders). Third, to investigate the relation between number of ratings and student evaluations, we computed correlations (Pearson's ρ) using all professors with at least 25 ratings. We performed these analyses to investigate whether professors with low or high student evaluations were more likely to motivate a greater number of ratings. As shown in Figure 1, we found statistically significant results in one of the four schools. However, as shown by R^2, the variance was small (range = 0%–1.9%) even in the isolated instance of a statistical difference.

Discussion

Our purpose in this study was to investigate perceived physical attractiveness and student evaluations of college professors by using data obtained from the Web site www.ratemyprofessors.com (i.e., a naturally occurring database of concurrent ratings). Results indicated that professors perceived as attractive received higher student evaluations than did nonattractive controls that were matched for both department and gender. In real numbers, professors perceived as attractive scored about 0.8 of a point higher on a 5-point scale (see Table 4). We interpret this difference as a practically meaningful result because professors perceived as attractive move from slightly higher than average on the 5-point scale (i.e., an okay professor) to above-average ratings (i.e., a good professor). With institutionally sponsored student evaluations, moving into the above-average category is often the difference on such important decisions for professors as promotion, tenure, and salary increases (Millea & Grimes, 2002; Williams & Ceci, 1997). Furthermore, results from this study

TABLE 5. Student Evaluations Controlled for Department and Separated by Gender

	Attractive		Nonattractive control				
School	*M*	*SD*	*M*	*SD*	*df*	*t*	Cohen's *d*
			Male professor				
Grand Valley State	4.18	0.48	3.44	0.82	104	5.73**	1.11
University of Delaware	4.08	0.86	3.54	0.72	40	2.20*	0.68
San Diego State	4.14	0.63	3.20	0.88	72	5.27**	1.23
James Madison	4.33	0.44	3.35	0.95	32	3.85**	1.32
			Female professor				
Grand Valley State	4.28	0.44	3.29	0.78	48	5.51**	1.56
University of Delaware	4.14	0.63	3.36	0.98	46	3.29*	0.95
San Diego State	4.11	0.40	3.59	0.59	30	2.90*	1.03
James Madison	4.61	0.36	3.39	0.65	12	4.36**	2.33

*$p < .05$. **$p < .001$.

were not an isolated finding, but were consistent across four separate universities. Perhaps the most interesting aspect of the results is the range of student evaluations. Ratings for professors perceived as nonattractive ranged from very low to extremely high (i.e., the full spectrum of student evaluations). However, ratings for professors perceived as attractive rarely dropped below an average score (only 6 out of 211 scored below an average rating of 3 on a 5-point scale).

Also, our overall results are consistent with a recent experimental investigation performed by Hamermesh and Parker (in press) in which the authors found a strong influence of perceived attractiveness on real end-of-semester evaluations. Klahr and Simon (2001) have advocated complementary approaches (both experimental and naturalistic) in the process of scientific discovery to provide convergent evidence. However, in contrast to the Hamermesh and Parker study in which the authors found a larger impact of perceived attractiveness for male professors, we found no evidence of gender differences because both male and female professors perceived as attractive received relatively equivalent ratings. Further research into the potential impact of gender differences and perceived attractiveness on student evaluations is warranted to determine the discrepancy between results.

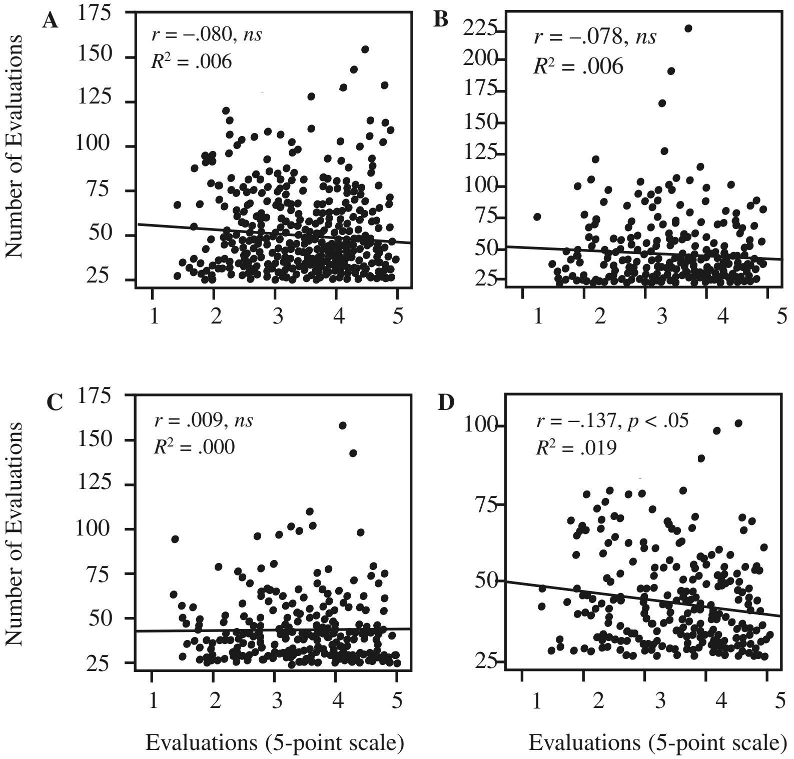

FIGURE 1. Scatterplots (with regression lines) of student evaluations and number of ratings (A = Grand Valley State, B = University of Delaware, C = San Diego State, D = James Madison University).

Of course, there are many potential limitations that could affect the overall validity of this study because we obtained data from a naturally occurring database. The most significant limitations are the lack of knowledge of the participants providing the ratings of professors and the potential for multiple ratings. There is no way to verify who provided the ratings because anyone could potentially contribute to the data. Likewise, the potential for multiple ratings is problematic because a single rater could artificially inflate or deflate a professor's overall rating. Despite the anonymous input, the basic characteristics of the ratings can be described. First, professors (with at least 25 ratings) had an average student evaluation of about 3.5 (see Table 1) indicating that most professors sampled were rated above average. Second, Figure 1 shows that the ratings are widely dispersed and not just clustered at the extremes on the 5-point student evaluation scale, indicating a wide distribution of input that is not solely targeted at evaluating professors rated as very poor (i.e., motivated to "slam" professors) and outstanding (i.e., motivated to praise professors) or both. Because students have a long history of disseminating and sharing information about professors (Williams & Ceci, 1997), it may be that the ratings, as indicated by the large number of inputs at these institutions (see Table 1), are most often used by students to communicate information.

It is important to note that although this study indicates that professors perceived as physically attractive receive higher student evaluations, our results should not be viewed in any way as intended to establish a causal link. Our results merely (a) add to the sparse current literature, (b) evaluate the potential for a practical data set analysis in contributing to the literature, (c) provide a complementary approach with ecological validity, and (d) lead to further research questions that can be evaluated using more rigorous experimental designs (as opposed to the naturalistic data collection used in this study).

It would be interesting for future research to determine the consistency of the initial perceptions of attractiveness (e.g., first day of classes) with attractiveness ratings at the end of the semester. As previously mentioned, multiple inputs beyond the actual physical appearance of the target and individual preferences of the perceiver contribute to the evaluation of physical attractiveness, which may vary with time (Eagly et al., 1991; Hosoda et al., 2003). Also, different levels of initial attractiveness (nonattractive, somewhat nonattractive, neutral, attractive, very attractive) may be more or less stable across time. Future researchers should attempt to establish the stability of the perception of attractiveness across time in the college classroom. Future researchers should also investigate whether evaluations of professors performed by peers, department chairs, and deans are also influenced by perceived attractiveness. Finally, future researchers should attempt to establish the validity of the student ratings at www.ratemyprofessors.com compared with those of expert evaluators and real in-class student evaluations.

REFERENCES

Abbey, A., Cozzarelli, C., McLaughlin, K., & Harnish, R. J. (1987). The effects of clothing and dyad sex composition on perceptions of sexual intent: Do women and men evaluate these cues differently? *Journal of Applied Social Psychology, 17,* 108–126.

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,* 431–441.

Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles, 43,* 407–417.

Benson, P. L., Karabenick, S. A., & Lerner, R. M. (1976). Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help. *Journal of Experimental Social Psychology, 12,* 409–415.

Bloch, P. H., & Richins, M. L. (1992). You look "mahvelous": The pursuit of beauty and the marketing concept. *Psychology & Marketing, 9,* 3–15.

Buck, S., & Tiene, D. (1989). The impact of physical attractiveness, gender and teaching philosophy on teacher evaluations. *Journal of Educational Research, 82,* 172–177.

Buckley, H. M. (1983). Perception of physical attractiveness as manipulated by dress: Subject versus independent judges. *The Journal of Psychology, 114,* 243–248.

Budesheim, T. L., & DePaola, S. J. (1994). Beauty or the beast? The effects of appearance, personality, and issue information on evaluations of political candidates. *Personality and Social Psychology Bulletin, 20,* 339–348.

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology, 84,* 563–572.

Cassidy, L., & Hurrell, R. M. (1995). The influence of victim's attire on adolescents' judgments of date rape. *Adolescence, 30,* 319–323.

Chaiken, S. (1979). Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology, 37,* 1387–1397.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but . . .: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110,* 109–128.

Freeman, H. R. (1994). Student evaluations of college instructors: Effects of type of course taught, instructor gender and gender role, and student gender. *Journal of Educational Psychology, 86*, 627–630.

Frieze, I. H., Olson, J. E., & Russell, J. (1991). Attractiveness and income for men and women in management. *Journal of Applied Social Psychology, 21*, 1039–1057.

Goebel, B., & Cashen, V. (1979). Age, sex and attractiveness as factors in student ratings of teachers: A developmental study. *Journal of Educational Psychology, 71,* 646–653.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209–1217.

Gross, A. E., & Crofton, C. (1977). What is good is beautiful. *Sociometry, 40,* 85–90.

Hamermesh, D. S., & Biddle, J. E. (1994). Beauty and the labor market. *American Economic Review, 84,* 1174–1194.

Hamermesh, D. S., & Parker, A. M. (in press). Beauty in the classroom: Instructors'pulchritude and putative pedagogical productivity. *Economics of Education Review*.

Horton, R. S. (2003). Similarity and attractiveness in social perception: Differentiating between biases for the self and the beautiful. *Self & Identity, 2,* 137–152.

Hosoda, M., Stone-Romero, E. F., & Coats, G. (2003). The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology, 56,* 431–462.

Johnson, D. J., & Rusbult, C. E. (1989). Resisting temptation: Devaluation of alternative partners as a means of maintaining commitment in close relationships. *Journal of Personality and Social Psychology, 57,* 967–980.

Johnson, K. P., & Workman, J. E. (1992). Clothing and attributions concerning sexual harassment. *Home Economics Research Journal, 21,* 160–172.

Klahr, D., & Simon, H. A. (2001). What have psychologists (and others) discovered about the process of scientific discovery? *Current Directions in Psychological Science, 10,* 75–79.

Klentz, B., Beaman, A. L., Mapelli, S. D., & Ullrich, J. R. (1987). Perceived physical attractiveness of supporters and nonsupporters of the women's movement: An attitude-similarity-mediated error (AS-ME). *Personality and Social Psychology Bulletin, 13,* 513–523.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A. Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126,* 390–423.

Mack, D., & Rainey, D. (1990). Female applicants' grooming and personnel selection. *Journal of Social Behavior and Personality, 5,* 399–407.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52,* 1187–1197.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92,* 202–228.

Mazzella, R., & Feingold, A. (1994). The effects of physical attractiveness, race, socioeconomic status and gender of defendants and victims on judgments of mock jurors: A meta-analysis. *Journal of Applied Social Psychology, 24,* 1315–1344.

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

Millea, M., & Grimes, P. W. (2002). Grade expectations and student evaluation of teaching. *College Student Journal, 36,* 582–590.

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology, 85,* 1035–1048.

Myers, D. G. (2005). Social psychology (8th ed.). New York: McGraw-Hill.

Naftulin, D., Ware, J., & Donnelly, F. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48,* 630–635.

Peterson's. (2004). Four-year colleges (35th ed.). Princeton, NJ: Author.

Radmacher, S. A., & Martin, D. J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *The Journal of Psychology, 135,* 259–268.

Riniolo, T. C., & Schmidt, L. A. (2000). Searching for reliable relationships with statistics packages: An empirical example of the potential problems. *The Journal of Psychology, 134,* 143–151.

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgments of physically attractive students: A review. *Review of Educational Research, 62,* 413–426.

Sigelman, C. K., Thomas, D. B., Sigelman, L., & Ribich, F. D. (1986). Gender, physical attractiveness, and electability: An experimental investigation of voter biases. *Journal of Applied Social Psychology, 16,* 229–248.

Simpson, J. A., Gangestad, S. W., & Lerma, M. (1990). Perception of physical attractiveness: Mechanisms involved in the maintenance of romantic relationships. *Journal of Personality and Social Psychology, 59,* 1192–1201.

Stewart, J. E. (1980). Defendant's attractiveness as a factor in the outcome of criminal trials: An observational study. *Journal of Applied Social Psychology, 10,* 348–361.

Ware, J., & Williams, R. (1975). The Dr. Fox effect: A study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education, 40,* 149–156.

Waters, M., Kemp, E., & Pucci, A. (1988). High and low faculty evaluations: Descriptions by students. *Teaching of Psychology, 15,* 203–204.

Williams, W. M., & Ceci, S. J. (1997). "How'm I doing?" *Change, 29,* 13–24.

Williamson, S., & Hewitt, J. (1986). Attire, sexual allure, and attractiveness. *Perceptual & Motor Skills, 63,* 981–982.

Workman, J. E., & Freeburg, E. W. (1999). An examination of date rape, victim dress, and perceiver variables within the context of attribution theory. *Sex Roles, 41,* 261–277.

Workman, J. E., & Orr, R. L. (1996). Clothing, sex of subject, and rape myth acceptance as factors affecting attributions about an incident of acquaintance rape. *Clothing & Textiles Research Journal, 14,* 276–284.

Manuscript received August 8, 2004
Revisions accepted for publication March 8, 2005

Quality Assurance in Education

Higher grades = higher evaluations: impression management of students
Philip R. OldsD. Larry Crumbley

Article information:

To cite this document:
Philip R. OldsD. Larry Crumbley, (2003),"Higher grades = higher evaluations: impression management of students", Quality Assurance in Education, Vol. 11 Iss 3 pp. 172 - 177
Permanent link to this document:
http://dx.doi.org/10.1108/09684880310495924

Downloaded on: 18 February 2016, At: 11:28 (PT)
References: this document contains references to 26 other documents.
To copy this document: permissions@emeraldinsight.com
The fulltext of this document has been downloaded 1073 times since 2006*

Higher grades = higher evaluations: impression management of students

Philip R. Olds and
D. Larry Crumbley

The authors

Philip R. Olds is Associate Professor at Virginia Commonwealth University, Richmond, Virginia, USA.
D. Larry Crumbley is KPMG Endowed Professor in the Department of Accounting, Louisiana State University, Baton Rouge, Louisiana, USA.

Keywords

Students, Examinations, Teachers

Abstract

This study investigates the effects on end-of-the semester evaluations of the instructor resulting from grade inflation by the administration of a different number of mid-semester tests to four different classes of the first course of accounting. Students in two classes received six mid-semester examinations, while the other two classes received three. All classes were given a cumulative final examination. Giving six examinations rather than three allows a student to memorize less of the course material before each examination, resulting in higher overall grades. Analysis of the data revealed that students receiving six mid-semester examinations scored higher than those receiving three. These differences were statistically significant at the 0.1 level. Students' end-of-the semester evaluations of the fairness of grading, quality of the instructor and the quality of the course were consistently more positive in the class that received six mid-semester examinations. Higher grades did result in higher evaluations. Two of these comparisons were statistically significant at the 0.01 level; the third was significant at the 0.5 level. The benefits of administering six mid-semester examinations exceeded the additional effort required.

Electronic access

The Emerald Research Register for this journal is available at
http://www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/0968-4883.htm

Quality Assurance in Education
Volume 11 · Number 3 · 2003 · pp. 172-177
© MCB UP Limited · ISSN 0968-4883
DOI 10.1108/09684880310495924

Administrator reliance upon student evaluation of teaching (SET) surveys to evaluate teaching effectiveness is an important and sensitive issue facing today's college-level faculty and administrators. There is growing controversy in the literature regarding the use of these instruments, as they play a vital role in the promotion, tenure, and merit process (Wallace and Wallace, 1998). Currently, SET instruments are used by more than 94 percent of accounting departments within schools of business, eight percentage points higher than the national usage (Calderon *et al.*, 1994).

Grading impacts evaluations

With the growing use of SET information, research has increasingly questioned the validity of these surveys as an indicator of instructor effectiveness (Bauer, 1996; Crumbley, 1995; Ellis, 1985). There is a strong positive correlation between grades received and overall course rating ($n = 468$, $r = 0.42$, $p < 0.0001$) (Williams and Ceci, 1997; Greenwald and Gillmore, 1997a). Greenwald and Gillmore (1997b) found that changing from giving grades one standard deviation below the university mean to one standard deviation above the university mean should produce a one standard deviation improvement in one's percentile rank in the university's student ratings (i.e. movement from the university's 31st percentile of instructors to the 69th percentile).

There is growing evidence in the literature that overemphasis on the numerical results of these survey instruments may be contributing to an erosion of quality of teaching and scholarship, to a lower level of respect for teachers, and to weakening of faculty positions (Haskell, 1997; Sacks, 1996). Registrars at even top schools believe that grades have accelerated faster than student talent levels. At Bucknell University, 80 percent of all grades given are As and Bs compared with 50 percent in the 1960s (Bulkeley, 1997).

Impression management

In a performance measurement system judged by student evaluations, classroom behavior and motives of some teachers may be partly explained through an impression management theory (also known as self-representation theory). As noted by Rosenfeld *et al.* (1994), "impression management refers to the many ways by which individuals attempt to control the impressions others have of them: their behavior, motivations, morality, and a host of person attributes." Schneider (1969) and Swann (1987) note that most individuals desire to be viewed in a favorable manner by others, and they construct a favorable image of themselves in order to maximize rewards, maintain their self-esteem, and create a desire self-identity.

Human nature suggests that, if you are in the position to evaluate an individual's work and do not provide a superior evaluation (e.g. an instructor giving a student low grades), such an individual may not evaluate you highly on an anonymous questionnaire. Centra and Creech (1976) report a moderately strong, statistically significant relationship between student grade expectations and the student rating of instructor effectiveness. Students expecting an "A" grade evaluated instructor effectiveness with a mean of 3.95, while those students expecting a "D" grade gave a mean rating of 3.02. An instructor's grading policy and course rigor may be significant factors in determining student responses on instructor evaluations. Many instructors may believe that Newton's (1988) leniency hypothesis is valid and take corrective actions to improve their evaluations by reducing coursework and increasing their grades. Ryan *et al.* (1980) note that at least one-third of their survey respondents indicated they have substantially decreased their grading standards and level of course difficulty. Bures *et al.* (1990) found that only 20.4 percent of 559 accounting professors agreed with the statement that SETs are indicative of an instructor's teaching and should be used directly in calculating annual salary increases.

If an instructor can choose teaching styles, grading difficulty, and course content, he or she may prefer the choices that are expected to result in higher SET scores. According to Medley (1979), "if teachers know the criteria on which decisions affecting their careers are based, they will meet the criteria if it is humanly possible to do so." Worthington and Wong (1979) argue that "as an instructor inflates grades, he or she will be much more likely to receive positive evaluations." Many SET enhancement choices have the potential to be dysfunctional or anti-learning, resulting in grade inflation, course work deflation, and "pander pollution" behavior. Pander pollution may be defined as purposeful intervention by an instructor inside and outside the classroom with the intention of increasing SET scores, which is counter-productive to the learning process. Increasing use of the SET has the potential for professors to engage in pander pollution in an attempt to enhance their SET scores[1].

Empirical evidence from the first course in accounting

The purpose of this study was to determine whether students' perceptions of the performance of an instructor are affected based on whether they receive six tests or three tests. To accomplish this research six tests were administrated to students in two classes and three tests to students in two other classes. These four classes were taught in two different semesters. Two were taught in the Fall of 1995 and two were taught in the Fall of 1996. During each semester, one class received three tests and the other received six. Students' test scores and evaluations of the instructor were then compared.

Classroom environment

Several factors were controlled as much as possible to ensure that if any differences in examination performance were discovered, these differences would be the result of examination frequency. The factors controlled included:

(1) All classes met on Mondays, Wednesdays, and Fridays for 50 minutes.
(2) During the Fall semester of 1995 the class that received three tests met at 11:00 am; the class that received six tests met at noon. During the Fall semester of 1996 the class that received three tests met at noon; the class that received six tests met at 11:00 am.
(3) The rooms in which the classes met were in the same building and in similar rooms. For example, both rooms were in the interior of the building.
(4) All classes were taught by the same instructor.
(5) Lectures in all classes were given using outlines generated with power point projected on to screen in color via a LCD panel system.

(6) Students in all classes were given printed copies of the PowerPoint outlines.
(7) The instructor had taught the course several times previously using the same PowerPoint outlines and projection equipment. Thus, there should not have been a "learning-curve effect" for the instructor.
(8) The same end-of-chapter problems were reviewed in each class.
(9) Other than the number of examinations, the grading policies of all classes were the same. For example, an attendance grade constituted approximately 6.5 percent of the course grade in each class. Also, in each class students were allowed to earn "extra credit" by turning-in homework that were collected on a random basis. The same number of homework problems was collected in each class and these problems were "weighted" to have the same potential effect on a student's grade in the course.
(10) The classes were of very similar sizes. For example, the class with the smallest number of students that took the test(s) over the first four chapters had 39 students and the largest was 45. (The number of students taking each test in each class is shown in Table I.)

Construction and administration of examinations

Eleven chapters of the textbook were covered in the course. In the classes that received six tests, an examination consisting of 12 selected-response questions was given after every two chapters, except for the last examination before the cumulative final. The last mid-semester examination consisted of six questions over the last chapter of the textbook.

The selected-response questions used on tests were a mixture of three different types: traditional multiple-choice, true-false, and "articulation" questions. The articulation questions were developed by the instructor to stress the inter-relationships among the balancesheet, income statement, and statement of cash flows. An example of each of the three types of questions is included in the Appendix. For each test, two versions (the same questions arranged in a different order) were prepared to reduce the possibility of students copying from one another. On each version questions were grouped by "type", but were randomly arranged within each group to reduce any effects of sequencing.

In the classes that received three tests, an examination consisting of 24 selected-response-type questions was administered after every four chapters, except for the last examination before the cumulative final. The last mid-semester examination consisted of 18 questions over the last three chapters of the textbook. The examinations administered in these classes were constructed by combining questions from the corresponding tests given in the six tests classes. For example, tests 1 and 2 from six tests classes were combined to create test 1 for the three tests classes. Questions for the three test version were also arranged into groups of similar type questions: articulation, multiple true-false, and multiple choice. However, within each group, the questions were arranged in the same sequence as they appeared on the six test versions. Again, two "scrambled' versions of each test were prepared to reduce cheating. All tests were timed so that students being tested over four chapters were given double the amount of time per tests that students who were tested over only two chapters were given.

Table I Comparisons of test scores of classes receiving six tests versus three tests

	H1A test(s) on first four chapters		*H1B* test(s) on second four chapters		*H1C* test(s) on last three chapters		*H2* cumulative final		*H3* total tests points	
Version	6 tests	3 tests	6 tests	3 tests	6 tests	3 tests	6 tests	3 tests	6 tests	3 tests
***Students*:**										
1995	39	41	37	40	34	37	35	37	34	37
1996	43	45	43	40	43	37	43	37	42	37
Total	82	86	80	80	77	74	78	74	76	74
Mean %	80.1	73.2	64.9	59.1	63.7	56.5	74.4	71.2	71.0	65.9
Std. dev.	12.47	15.72	13.07	12.08	13.58	14.60	13.39	14.52	10.46	11.87
***P*-value**	0.002		0.004		0.002		0.163		0.006	

Hypotheses tested

Four separate research hypotheses were tested:

H1. Students' scores on mid-semester tests will not be affected by the frequency with which examinations are administered.

H2. Students' scores on the cumulative final examination will not be affected by the frequency with which examinations are administered.

H3. The sum of students' scores on all tests given during the course will not be affected by the frequency with which examinations are administered.

H4. Students' perception of the fairness of the instructor's grading methods, quality of the instructor compared with other instructors, or quality of the course compared to other courses will not be affected by the frequency with which examinations are administered.

These hypotheses were tested by using *t*-tests to compare students' test scores and end-of-the-semester evaluations of the instructor. A total of eight tests were performed: three to test *H1*, one each for *H2* and *H3*, and three to examine *H4*.

Results of the data analysis

Table I shows the results of the analysis of *H1-3*. The differences in test scores on all the examinations except the cumulative final were significant at the 0.1 level. In each case, the scores of the classes that received six tests were higher than those of the classes receiving three tests. However, the difference was not statistically significant at even the 0.10 level. Students who took six examinations had a higher cumulative score than those who took only three, and this difference was significant at the 0.01 level. Thus, *H1* and *H3* were rejected, while *H2* could not be rejected.

These results are generally consistent with those obtained in other studies. A similar experiment with students enrolled in a tax accounting course (Murphy and Stanga, 1994) found that students receiving six mid-semester examinations received higher scores than those receiving only three, and these differences were sometimes, but not always, statistically significant. A meta-analysis of 31 prior studies on the effects of testing frequency found that students generally perform better if more tests are administered than if only a few are given (Crooks, 1988, pp. 448-9). However, in the studies reviewed, "few" sometimes meant that only a final examination was given. The conclusion of that study was that instructors should give one or two mid-semester tests, but little additional benefit results from giving more. This conclusion is similar to the recommendation that was made by Jacobs and Chase (1992, p. 30). They suggested that two or three mid-semester tests be given.

However, there is a greater benefit to an instructor for more frequent testing beyond the students' grades in the course. This study, like Murphy and Stanga (1994) and Fulkerson and Martin (1981), found that students' end-of-the-semester evaluations of instructors are more positive in classes that receive more tests. Students may "sell" higher evaluations for higher grades. Thus, an instructor can manage the impressions that students have of the instructor by increasing the number of examinations and increasing the overall grade point average.

Table II presents the results of analyzing students' responses to the three statements on the evaluation form used in the study. These statements were 1: "The instructor's method of grading was fair," and 2: "Compared with other instructors I have had at this institution, this one was fair." Students responded on a scale of one to five, where five was the most positive response and one was the most negative. As the data show, the students that received six tests gave more positive responses to each question than students who received three tests. All of these comparisons were significant at the 0.01 level, so *H4* can also be rejected.

Given the importance that many schools place on students' evaluations of instructors, and considering that more frequent testing certainly does not seem to diminish learning and may improve it in some cases, improved evaluations alone might provide sufficient justification for administering tests more frequently. However, another benefit has been established by prior research. Students who receive more frequent testing have less test anxiety (Fulkerson and Martin, 1981; Peckham and Roe, 1977). Furthermore, the Fulkerson and Martin (1981) study found that the more test anxiety a student has, the more likely his or her grade will be improved by more frequent testing. Thus, the more frequent testing will improve one's evaluations without appearing to be engaged in unethical impression management.

Table II Comparisons of students' end-of-semester course evaluations

	H4A Fairness of grading		*H4B* Quality of instructor		*H4C* Quality of course	
Class	**Many**	**Few**	**Many**	**Few**	**Many**	**Few**
No. students						
1995	27	34	37	34	27	34
1996	32	30	32	30	32	30
Total	59	64	59	64	59	64
Mean	4.2	3.8	4.0	3.5	3.6	3.0
Std. dev.	0.845	1.142	0.833	0.791	0.784	0.984
***P*-value**	0.015		0.000		0.001	

Conclusions

Giving students numerous mid-semester tests may provide several benefits to the students and the instructor at little additional cost. These benefits include higher scores for some students and better evaluations for the instructor. An instructor should certainly try this impression management technique when faced with the heavy use of student evaluation data by administrators. This impression management technique does not clearly fall into the so-called "pander pollution area."

Note

[1] Laws highly regulate financial statements to reduce income manipulation and opportunistic behavior; yet there is little regulation of grade inflation and SET manipulation. Most administrators blindly accept them as truth. Instructors have a high incentive to manage SET, even more so than managers have the incentive to enhance earnings. See, for example, Dechow *et al.* (1995).

References

Bauer, H.H. (1996), "The new generations: students who don't study", *The Technological Society at Risk Symposium*, Orlando, FL, September 10, pp. 1-37.

Bulkeley, W.M. (1997), "Would tax plan further inflate college grades?", *Wall Street Journal*, April 22, pp. B1 and B7.

Bures, A.L., DeRidder, J.J. and Tong, H-M. (1990), "An empirical study of accounting faculty evaluation systems", *The Accounting Educators' Journal*, Summer, pp. 68-76.

Calderon, T.G., Green, B.P. and Reider, B.P. (1994), "Extent of use of multiple information sources in accessing accounting faculty teaching performance", working paper, May, pp. 1-22.

Centra, J.A. and Creech, F.R. (1976), "The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness", *SIR Report*, No. 4, Educational Testing Service, Princeton, NJ, pp. 24-7.

Crooks, T.J. (1988), "The impact of classroom evaluation practices on students", *Educational Research*, Winter, pp. 438-81.

Crumbley, D.L. (1995), "The dysfunctional atmosphere of higher education: games professors play", *Accounting Perspectives*, Spring, pp. 67-76.

Dechow, P.M., Sloan, R.G. and Sweeney, A.P. (1995), "Detecting earnings management", *The Accounting Review*, April, pp. 193-225.

Ellis, R. (1985), "Ratings of teachers by their students should be used wisely – or not at all", *The Chronicle of Higher Education*, November 20, p. 88.

Fulkerson, F.E. and Martin, G. (1981), "Effects of exam frequency on student performance, evaluation of instructor, and test [*sic*] anxiety", *Teaching of Psychology*, April, pp. 90-3.

Greenwald, A.G. and Gillmore, G.M. (1997a), "Grading leniency is a removable contaminant of student ratings", *Journal of Educational Psychology*, Vol. 89 No. 4, pp. 1209-16.

Greenwald, A.G. and Gillmore, G.M. (1997b), "No pain, no gain? The importance of measuring course workload in student ratings of instructions", *Journal of Educational Psychology*, Vol. 89 No. 4, pp. 743-51.

Haskell, R.E. (1997), "Academic freedom, tenure, and student evaluation of faculty: galloping polls in the twentieth century", *Education Policy Analysis Archives*, Vol. 5 No. 6, pp. 1-32.

Jacobs, L.C. and Chase, C.I. (1992), *Developing and Using Tests Effectively*, Jossey-Bass, San Francisco, CA.

Medley, D.M. (1979), "The effectiveness of teachers", in Peterson, P.O. and Walberg, H.J. (Eds), *Research on Teaching: Concepts, Findings and Implications*, McCutchan Publishing, Berkeley, CA, pp. 11-27.

Murphy, D.P. and Stanga, K.G. (1994), "The effects of frequent testing in an income tax course: an experiment", *Journal of Accounting Education*, Vol. 6. pp. 27-41.

Newton, J.D. (1988), "Using student evaluation of teaching in administration control: the validity problem", *Journal of Accounting Education*, Vol. 6, pp. 1-14.

Peckham, P.D. and Roe, M.D. (1977), "The effects of frequent testing", *Journal of Research and Development in Education*, No. 3, pp. 40-50.

Rosenfeld, P., Giacalone, R.A. and Riordan, C.A. (1994), "Impression management theory and diversity", *American Behavioral Scientist*, March, Vol. 12 No. 4, pp. 601-20.

Ryan, J.J., Anderson, J.A. and Birchler, A.B. (1980), "Student evaluation: the faculty responds", *Research in Higher Education*, Vol. 12 No. 4, pp. 317-33.

Sacks, P. (1996), *Generation X Goes to College*, Open Court, Chicago, IL.

Schneider, D.J. (1969), "Tactical self-preservation after success and failure", *Journal of Personality and Social Psychology*, No. 13, pp. 262-8.

Swann, W.B. (1987), "Identity negotiations: where two roads meet", *Journal of Personality and Social Psychology*, Vol. 53, pp. 1038-51.

Wallace, J.J. and Wallace, W.A. (1998), "Why the costs of student evaluations have long since exceeded their value", *Issues in Accounting Education*, May, Vol. 13 No. 2, May, pp. 443-7.

Williams, W.M. and Ceci, S.J. (1997), "How'm I doing", *Change*, September/October, pp. 13-23.

Worthington, A.G. and Wong, P.T.P. (1979), "Effects of earned and assigned grades on student evaluations of an instructor", *Journal of Educational Psychology*, Vol. 71 No. 6, pp. 764-75.

Appendix. Examples of each type of question used on examinations

Articulation questions

For each situation below, indicate its effects on the accounting elements shown on the accompanying "chart." Use the following letters to indicate your answer (you do not need to enter amounts.)

Increase = **I**; Decrease = **B**; No effect = **N**

1. Joyce Co. provided $2,000 of services on account.

Assets	Liabilities	Equity	Revenues	Expenses	Net income	Cash

Multiple true-false questions

(Use the answer sheet provided on the cover page.)

6. Indicate if each of the following statements about assets is true or false:
   A. Sales made on account have no immediate effect on assets.
   B. Assets are the same thing as economic resources to a business entity.
   C. Decreases in retained earnings are always offset by decreases in assets.
   D. As long as a company has more assets than liabilities, it will be able to pay its bills.
   E. Assets are recorded on the balance-sheet at their fair market value (i.e. what they would be worth if sold).

Multiple true-false questions

(Use the answer sheet provided on the cover page.)

9. Which of the following would be included in the "Cash flow from investing activities" section of the Statement of Cash Flows:
   A. Borrowed $10,000 from local bank.
   B. Purchased land with cash.

This article has been cited by:

1. Kelly R. Schutz, Brent M. Drake, Janet Lessner, Gail F. Hughes. 2015. A Comparison of Community College Full-Time and Adjunct Faculties' Perceptions of Factors Associated With Grade Inflation. *The Journal of Continuing Higher Education* **63**, 180-192. [CrossRef]
2. Donald Larry Crumbley, Ronald E. Flinn, Kenneth J. Reichelt. 2010. What is Ethical About Grade Inflation and Coursework Deflation?. *Journal of Academic Ethics* **8**, 187-197. [CrossRef]
3. Rod Gapp, Ron Fisher. 2006. Achieving excellence through innovative approaches to student involvement in course evaluation within the tertiary education sector. *Quality Assurance in Education* **14**:2, 156-166. [Abstract] [Full Text] [PDF]
4. Joseph K. Cavanaugh. 2006. What did you get? A faculty grade comparison. *Quality Assurance in Education* **14**:2, 179-186. [Abstract] [Full Text] [PDF]

Cognitive Dissonance or Revenge? Student Grades and Course Evaluations

Trent W. Maurer
Department of Hospitality, Tourism, and Family & Consumer Sciences
Georgia Southern University

I tested 2 competing theories to explain the connection between students' expected grades and ratings of instructors: cognitive dissonance and revenge. Cognitive dissonance theory holds that students who expect poor grades rate instructors poorly to minimize ego threat whereas the revenge theory holds that students rate instructors poorly in an attempt to punish them. I tested both theories via an experimental manipulation of the perceived ability to punish instructors through course evaluations. Results indicated that student ratings appear unrelated to the ability to punish instructors, thus supporting cognitive dissonance theory. Alternative interpretations of the data suggest further research is warranted.

Given the reliance of many university administrators on student evaluations of teaching as a method for evaluating faculty job performance, it is not surprising that many faculty are concerned about the validity of student evaluations (Academic Job Forum, 2005). Although prior research has identified numerous factors as potentially biasing variables (for an overview, see Marsh & Dunkin, 1992), arguably the most controversial biasing factor is students' expected grade (Ginexi, 2003). The literature has repeatedly documented a significant relation between expected grade and student ratings (Marsh & Dunkin, 1992; Wachtel, 1998), and there is some evidence that this relation is causal. Salmons (1993) reported that when comparing student ratings from early in the semester with later in the semester, students who expected to receive an F at the end of the course lowered their ratings from the first evaluation, whereas students who expected to receive an A or B raised their ratings.

A common explanation for this relation is that students who are dissatisfied with their grades attempt to seek revenge against their instructors by rating them poorly, hoping that poor ratings will result in the instructors' termination or other negative consequences (Academic Job Forum, 2005). There are two problems with this explanation. First, there is extremely limited empirical evidence to support this interpretation. Although the influence of expected grade on student ratings is well documented, the nature of the relation is relatively unknown. Second, there is significant evidence that students are either unaware of the use of evaluations in making personnel decisions or do not believe that ratings will significantly influence personnel decisions (Chen & Hoshower, 2003; Marlin, 1987; Spencer & Schmelkin, 2002).

Alternative research has suggested cognitive dissonance as an explanation of the relation between expected grade and student ratings (Ginexi, 2003). That is, when students expect to receive a high grade but instead receive a low grade, they are confronted with a discrepancy that they must explain. They can either attribute the discrepancy to internal causes (e.g., failure to study, believing one is "stupid") or external causes (e.g., the instructor was unfair). Because an internal attribution would be threatening to the ego or self-esteem, students attempt to protect their self-image by locating the responsibility for the discrepancy externally and blaming the instructor. If the relation between expected grade and student ratings is driven by cognitive dissonance, one would anticipate that only the ratings of the instructor would be influenced by expected grade and that other elements of the course, such as the appropriateness of the textbook, would remain unaffected. This pattern is precisely what Ginexi (2003) reported.

However, it is also possible that the pattern of results Ginexi (2003) reported could be explained by the revenge theory (i.e., if students wanted to get even with an instructor for a poor grade, they would likely rate the instructor poorly, but not the unrelated elements of the course, such as the textbook, so that it would not be immediately obvious from the pattern of their responses that they were simply trying to get even). What is needed is a test of the two competing theories that specifically controls for the possibility that students may attempt to punish their instructors for low grades. Although an exact test is not present in the literature, a related investigation may inform this inquiry. Kasten and Young (1983) reported an experimental manipulation in which 77 graduate students completed a midterm evaluation form administered with one of three sets of instructions. Students in the first group read that the purpose of the evaluations was for personnel decisions, such as salary awards. Students in the second group read that the purpose of the evaluations was for course improvement and that administrators would not see them. Students in the third group read no attributional instructions. The authors' analyses indicated no significant differences between the groups in the ratings of the instructors, suggesting the cognitive dissonance theory to be correct.

Unfortunately, this investigation had several limitations that make it less than ideal for informing faculty about the nature of the expected grade–student rating relation. First, the authors used a relatively small sample that may have had insufficient power to detect a small but significant difference between conditions. Second, the participants were graduate students, but the majority of the literature and discourse on course evaluations focuses on undergraduates. Third, students completed the evaluations at midterm rather than at the end of course (to allow for the experimental manipulation), and prior research has documented that student evaluations at midterm may not be reliable (Salmons, 1993). Fourth, the instructions that contained the experimental manipulation were written, and students may not have read them (thus invalidating the manipulation). Fifth, the authors were unable to control for differences in instructor because different instructors taught the courses. Finally, the data from this study are nearly 25 years old, and a more modern investigation could incorporate the past quarter-century of research on course evaluations. For example, Kasten and Young (1983) did not even assess expected grade or test for a relation between expected grade and course evaluation (or an interaction among expected grade, condition, and course evaluation).

I attempted to address the limitations of Kasten and Young (1983) by collecting student ratings from several hundred undergraduate students over a 2-year period. The same instructor taught all of the students and administered oral instructions to manipulate their experimental condition. I hypothesized that there would be a significant positive association between expected grade and student ratings of instructor. Also, if the revenge theory were correct, then a significant Grade × Condition interaction would appear, such that students in the personnel decision condition who anticipated low grades would rate the instructor lower than students who anticipated the same grades in the course improvement condition.

Method

Students in 17 classes completed course evaluation forms. I taught the classes in the fall, spring, and summer semesters from 2003 to 2005, assigning one of the three conditions to each class each semester. The four summer classes completed evaluations in the control condition because it is not mandatory for faculty at this university to administer course evaluations in the summer and administrators do not use summer course evaluations in faculty evaluations. I assigned the fall and spring classes to the experimental conditions in a counterbalanced fashion (see Table 1) to control for any possible effect due to instructor improvement over time. The control or course improvement condition (C) instructed students to take the evaluations seriously even though university administrators would not see them and they would have no effect on decisions about promotion, tenure, hiring, or firing of the instructor. The first experimental or personnel decision condition (A) instructed students to take the evaluations seriously because university administrators would use them to make decisions about promotion, tenure, hiring, and firing of the instructor. The second experimental condition (B) instructed students only to take the evaluations seriously.

All evaluations were voluntary and anonymous, and students completed them in the last 2 weeks of class on a nonexam day. A total of 642 students completed evaluations. The evaluation forms did not collect demographic information, but approximately 90% of the students enrolled in these courses were women. The two questions on the evaluation used in this investigation were: "Overall, how would you rate this instructor?" with a scale ranging from 1 (*very poor*) to 5 (*very good*); and "What grade do you expect in this course?" with a scale ranging from 1 (*F*) to 5 (*A*).

Results

A univariate ANOVA tested the effect of expected grade, experimental condition, and the Grade × Condition interaction on student ratings of the instructor. Calculations of effect size used partial eta squared. The overall model was significant, $F(9, 632) = 4.79$, $p < .001$, $\eta_p^2 = .06$, with a significant main effect for expected grade, $F(3, 632) = 9.73$, $p < .001$, $\eta_p^2 = .04$. Post hoc analyses using Tukey's honestly significant difference revealed that students who expected to receive a D in the course rated the instructor lower than students who expected As, Bs, or Cs, and students who expected Cs rated the instructor

Table 1. Courses and Conditions Assigned

	2003–2004			2004–2005		
Course	Fall	Spring	Summer	Fall	Spring	Summer
Family development	A	B	C	B	A	C
Child development	—	—	C	—	—	C
Lifespan development	B	—	—	A	—	—
Research methods	—	B	—	B	A	—
Prenatal and infant development	A	B	—	B	A	—

Note. A = personnel condition; B = course improvement condition; C = control condition.

Table 2. Means by Group and Post Hoc Analyses

	Group Means				
Item	1	2	3	4	5
Expected grade	—	2.33_a	4.17_b	$4.37_{b,c}$	4.51_c
Experimental condition	4.25	3.91	4.37	—	—

Note. In each row, means with different subscripts are significantly different. For expected grade, Group 1 = F; Group 5 = A. For experimental condition, Group 1 = Condition A (personnel); Group 2 = B (course improvement); Group 3 = C (control).

lower than students who expected As. The effect for experimental condition was not significant, $F(2, 632) = 2.28$, $p = .10$, nor was the Grade × Condition interaction, $F(4, 632) = 1.77$, $p = .13$. Results appear in Table 2.

Discussion

The results of this investigation replicated and extended the work of Kasten and Young (1983) and appear to support the cognitive dissonance explanation for the connection between students' expected grades and their evaluations of instructors. That is, as hypothesized, a significant effect for expected grade on ratings of the instructor appeared, but contrary to the second hypothesis, no interaction appeared between experimental condition and student ratings. Students who expected lower grades did rate the instructor lower than students who reported higher grades, but the magnitude of this effect was not larger in the personnel decision condition or smaller in the control or course improvement condition.

Although these results may suggest that students do not rate instructors lower in an attempt to seek revenge for lower grades, I believe it would be overstating the results to interpret them as a complete rejection of the revenge theory. The absence of proof should not be taken as proof of absence. There are at least three alternative explanations for why a significant interaction between grade and condition did not emerge from the data. First, only 3 students reported that they expected to receive a D in the course, and all were in experimental condition B. No students reported that they expected to receive an F. (In some respects, this result is not surprising given that one would not expect students who are doing so poorly to be attending class regularly.) This distinction is important because at this university (like many others), students who receive less than a C in a major course must retake the course. Having to retake a course could be a major motivation (both financial and otherwise) to attempt to seek revenge against an instructor, but it was virtually impossible to assess in this sample. Without a substantially larger number of students who anticipated receiving less than a C in the course (and who are distributed equally across the conditions), it is impossible to conclusively rule out revenge as a student motivation. (However, it should be noted that at least the students who receive As, Bs, or Cs did not appear to be motivated by revenge in their evaluations of instructors, and these students accounted for over 99% of the sample.)

Second, many students either do not know or do not believe that administrators use evaluations in making personnel decisions (Chen & Hoshower, 2003; Marlin, 1987; Spencer & Schmelkin, 2002). Although the experimental manipulation attempted to address this problem, it is still possible that students continued to believe that administrators would not use their evaluations, and the experimental manipulation may not have influenced students. However, regardless of the success of the manipulation, the data suggest that students do not rate instructors out of a desire for revenge: If students did believe that evaluations could change things, then the manipulation was likely successful and the conclusion that revenge theory is unsupported stands. If students did not believe that evaluations could change things, one could alternatively interpret the data to mean that students will not seek revenge against their instructors even when given the opportunity to do so because they do not believe they can get revenge. In either case, the data suggest that revenge theory is unsupported (albeit for different reasons).

Third, due to structural constraints, it was possible only to administer conditions to entire classes rather than randomly to individual students in each class, and it was possible to administer the control condition only in the summer term, which prevented truly random assignment of experimental conditions across semesters and within classes. It is possible that the nature of summer courses (e.g., their shorter duration or a selection effect in the students who enroll in them) introduced some form of error into the experiment that masked a difference between conditions (although this possibility alone would not explain the lack of a significant difference between the two experimental conditions). However, Kasten and Young (1983) specifically called for this kind of class-based replication of their research to minimize within-group differences and more effectively test for differences across conditions. I encourage faculty at institutions that allow more freedom in manipulating course evaluation forms and administration procedures to attempt to replicate this research using random assignment of conditions at the individual level to determine if it may be possible to find support for revenge theory.

One reviewer suggested that the absence of support for revenge theory in this investigation would discourage others from pursuing future research exploring the theory as an alternative interpretation of the grade–evaluation relation. Although this concern is not surprising given the maxim, "Nobody publishes null results," I see these findings rather as a challenge to the proponents of revenge theory. Given the prevalent lay opinion frequently expressed among faculty that the revenge theory is the appropriate explanation for the grade–rating relation, a null finding in this research is important. Namely, the first large-scale testing of revenge theory was unable to substantiate its claim. As scientists, faculty should not use explanations of phenomena absent empirical support (however ego-soothing they may be). In fact, one could argue that clinging to a revenge theory explanation of poor course evaluations in the absence of any evidence for it is itself evidence of cognitive dissonance; faculty attempt to locate discrepant negative feedback about their teaching externally so as not to threaten to their own "good teacher" identity. The lack of support for revenge theory in this investigation should not discourage other researchers from further investigating the phenomenon; rather, it should encourage proponents of revenge theory to conduct further research to empirically substantiate their claims.

References

Academic Job Forum. (2005, June 17). *The Chronicle of Higher Education,* p. C4.

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education, 28,* 71–88.

Ginexi, E. M. (2003). General psychology course evaluations: Differential survey response by expected grade. *Teaching of Psychology, 30,* 248–251.

Kasten, K. L., & Young, I. P. (1983). Bias and the intended use of student evaluations of university faculty. *Instructional Science, 12,* 161–169.

Marlin, J. W., Jr. (1987). Student perception of end-of-course evaluations. *Journal of Higher Education, 58,* 704–716.

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, pp. 143–233). New York: Agathon.

Salmons, S. D. (1993). The relationship between students' grades and their evaluation of instructor performance. *Applied H.R.M. Research, 4,* 102–114.

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment and Evaluation in Higher Education, 27,* 397–409.

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

Note

Send correspondence to Trent W. Maurer, Department of Hospitality, Tourism, and Family & Consumer Sciences, P.O. Box 8021, Georgia Southern University, Statesboro, GA 30460; e-mail: tmaurer@georgiasouthern.edu.

Student Ratings

The Validity of Use

Wilbert J. McKeachie
University of Michigan

In this article, the author discusses the other articles in this Current Issues *section and concludes that all of the authors agree that student ratings are valid but that contextual variables such as grading leniency can affect the level of ratings. The authors disagree about the wisdom of applying statistical corrections for such contextual influences. This article argues that the problem lies neither in the ratings nor in the correction but rather in the lack of sophistication of personnel committees who use the ratings. Thus, more attention should be directed toward methods of ensuring more valid use.*

I chuckled with pleasure at some of the thrusts and counterthrusts as I read the preceding articles by Greenwald (1997, this issue), Marsh and Roche (1997, this issue), d'Apollonia and Abrami (1997, this issue), and Greenwald and Gillmore (1997, this issue) in this *Current Issues* section. Each article contains much good sense. My role, presumably, is to give an overview in terms of my experience as a researcher, a teacher, and an evaluator of teaching both for improvement and for personnel decisions. The articles in this section address three main issues: (a) How many dimensions of teaching should student rating forms report to personnel committees? (b) Are student ratings valid measures of teaching effectiveness? and (c) Are student ratings biased by variables other than teaching effectiveness, and if so, can these biases be controlled statistically? I shall briefly address each of these issues. Then I argue that the basic problem is not with the ratings but rather with the lack of sophistication of those using them for personnel purposes. I conclude with some observations and recommendations for research and practice.

How Many Dimensions of Teaching Should Student Rating Forms Report?

The answer to this question depends on what one wants to do with the ratings. Most people interested in improving teaching see the primary purpose of student ratings as providing feedback to teachers that will be helpful for improvement. General overall ratings provide little guidance. Murray (1983, 1997) has shown that specific behavioral items are most likely to result in improvement. Renaud and Murray (1997) have shown that actual behaviors of teachers as coded by observers covary with student ratings of the same behaviors and fall into dimensions corresponding fairly well to those of Marsh (1984). Marsh's demonstration of the validity of these factors is impressive. Grouping items by factors can reduce the "mental dazzle" of a long computer printout of many items and can increase the likelihood of improvement.

But what about reports to committees or administrators making personnel decisions? Such a committee must arrive at a single judgment of overall teaching effectiveness. If one grants that overall ratings of teaching effectiveness are based on a number of factors, should a score representing a weighted summary of the factors be represented (as Marsh and Roche [1997] argue), or should one simply use results of one or more overall ratings of teaching effectiveness (as contended by d'Apollonia and Abrami, 1997)? I would prefer student ratings of attainment of educational goals rather than either of these alternatives. Whatever score or scores are used, I agree with d'Apollonia and Abrami's conclusion: "We recommend that . . . only crude judgments of instructional effectiveness (exceptional, adequate, and unacceptable) [be made on the basis of student ratings]" (p. 1205).

The first reason for a simple three-category classification is that personnel committees do not need to make finer distinctions. The most critical decision requires only two categories—"promote" or "don't promote." Even decisions about merit increases require no more than a few categories, for example, "deserves a merit increase," "deserves an average pay increase," or "needs help to improve."

A second reason for endorsing d'Apollonia and Abrami's (1997) view is that effective teachers come in many shapes and sizes. Scriven (1981) has long argued that no ratings of teaching style (e.g., enthusiasm, organization, warmth) should be used, because teaching effectiveness can be achieved in many ways. Using characteristics that generally have positive correlations with effectiveness penalizes the teacher who is effective despite less than top scores on one or more of the dimensions

I thank Matt Kaplan, Diana Kardia, and James Kulik for their helpful comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Wilbert J. McKeachie, Department of Psychology, University of Michigan, 525 East University, Ann Arbor, MI 48109-1109. Electronic mail may be sent via Internet to billmck@umich.edu.

Vol. 52, No. 11, 1218–1225

usually associated with effectiveness. Judging an individual on the basis of characteristics, Scriven says, is just as unethical as judging an individual on the basis of race or gender.

A third problem with a profile of scores on dimensions is that faculty members and administrators have stereotypes about what good teaching involves. In most meetings to make decisions about promotions or merit salary increases, negative information is likely to be weighted more heavily than positive information. Thus, teachers who do not conform to the stereotype are likely to be judged to be ineffective despite other evidence of effectivenes. My colleagues and I have found evidence of this effect in our studies of the use of student ratings in promotion decisions (Lin, McKeachie, & Tucker, 1984; Salthouse, McKeachie, & Lin, 1978). If personnel committees sensibly use broad categories rather than attempting to interpret decimal-point differences, either a single score or a weighted combination of factor scores will provide comparable results.

Do Student Ratings Provide Valid Data About Teaching Effectiveness?

Evaluation for Improving Teaching

In the articles in this *Current Issues* section, there is little disagreement about the usefulness of student ratings for improvement of teaching (at least when student ratings are used with consultation or when ratings are given on specific behavioral characteristics). There are, however, two problems that detract from the usefulness of ratings for improvement.

The first problem involves students' conceptions of effective teaching. Many students prefer teaching that enables them to listen passively—teaching that organizes the subject matter for them and that prepares them well for tests. Unfortunately, most college teachers are not well trained in test construction. Even teachers who have development of students' thinking as a primary goal give examinations that primarily involve rote memory (McKeachie & Pintrich, 1991).

Cognitive and motivational research, however, points to better retention, thinking, and motivational effects when students are more actively involved in talking, writing, and doing (McKeachie, 1951; Murray & Lan, 1997). Thus, some teachers get high ratings for teaching in less than ideal ways.

The second problem is the negative effect of low ratings on teacher motivation. If a teacher is already anxious, then ratings that confirm the impression that students are bored or dissatisfied are not likely to increase the teacher's motivation and eagerness to enter the classroom and face the students.

A solution for both of these problems is better feedback. Marsh and Roche (1993) demonstrated that feedback targeted to specific problems identified by student ratings results in improvement. Murray and Smith (1989) found that items on specific teaching behaviors resulted in greater improvement than ratings on more general characteristics. In addition, research shows that student ratings are more helpful if they are discussed with a consultant or a peer (Aleamoni, 1978; Cohen, 1980; Marsh & Overall, 1979; McKeachie et al., 1980). Ideally, consultation should be only one feature of an academic culture in which colleagues discuss teaching and both teachers and students develop a sophisticated understanding of what is most helpful for lasting learning.

Evaluation for Promotion

But what about the use of student ratings for personnel decisions? Here again, the authors of the articles in this *Current Issues* section provide reassurance. All of the authors (and I join them) agree that student ratings are the single most valid source of data on teaching effectiveness. In fact, as Marsh and Roche (1997) point out, there is little evidence of the validity of any other sources of data.

However, student ratings are not perfectly correlated with student learning, even in the validity studies carried out in large courses with multiple sections. Many multisection courses use objective tests that assess factual knowledge. In these courses, students' ratings of teaching effectiveness are likely to reflect a relatively unsophisticated conception of effectiveness.

What is effective, however, is more complex. It depends on one's definition of the goals of teaching. If one believes that retention and later use of course concepts are important, mere presentation of the subject and testing for memory of facts is not likely to be effective. If one believes that outcomes such as skills for continued learning and critical thinking, motivation for lifelong learning, and changes in attitudes and values are important, it becomes clear that effective teaching must involve much more student talking, writing, and doing as well as evaluation methods that probe more deeply than most true–false or multiple-choice tests.

I agree with Marsh and Roche's (1997) statement that researchers need to provide validity data that go beyond recall of facts. Both Marsh and I have found student ratings to be valid with respect to other criteria, including motivational, attitudinal, and other goals of education (Marsh, 1984; McKeachie, Guetzkow, & Kelly, 1954; McKeachie, Lin, & Mann, 1971; McKeachie & Solomon, 1958).

The good news is that student ratings correlate positively with these indexes of teachers' effectiveness. The bad news is that teachers are not equally effective for all goals and all students. Cross (1958) found in one multisection study that there was a negative correlation between teachers' effectiveness as measured by the multiple-choice portion of the final examination and effectiveness as measured by the essay portion of the final examination. Hoyt and Cashin (1977) found that teaching behaviors associated with learning factual knowledge were different from those that help students develop problem-solving skills or self-understanding. Thus, a personnel committee needs to consider the relative importance of different educational goals when assessing teaching.

In addition to the need to look at other outcomes, researchers need to be aware of two additional problems with multisection studies. The first problem is that multisection courses are primarily first- and second-year courses; student ratings in these courses may have lower validity coefficients than in more advanced courses in which students have broader experience (and perhaps greater educational sophistication) as a basis for their ratings.

The second problem is that the achievement measure is common to all sections. Thus, what it assesses with respect to teaching is how well the teacher has prepared students for the test; it does not assess learning that goes beyond the test. And the test almost necessarily must be based on common material in the textbook. A classic study by Parsons (1957) found that students who simply studied the textbook without any classroom instruction did better on the final course examination than did those who had conventional classroom instruction that went beyond the textbook.

Isaacson, McKeachie, and Milholland (1963) found that the teaching assistants who were most effective had been rated by their peers as having broad cultural interests and knowledge. Good teachers often go well beyond the textbook. To get a valid measure of real teaching effectiveness, researchers need to measure not only what is taught in common but also educational gains that go beyond the minimum measured by a common examination. Students' papers, journals, and measures of motivation and attitude or other outcomes are needed.

Not only do good teachers go beyond the textbook, but their influence goes well beyond the geographical confines of the classroom. Most student rating forms and most faculty members' evaluations of teaching effectiveness focus almost completely on conventional classroom teaching. Clearly, much—very likely most—student learning occurs outside the classroom. Researchers need to get data on teachers' out-of-class contributions to education (d'Apollonia and Abrami's, 1997, "teacher as manager"). Student rating forms need to cue students to consider these aspects of teaching in their ratings.

Are Student Ratings Biased by Other Variables?

Greenwald and Gillmore (1997) are concerned about at least two sources of bias—class size and grading leniency. The concern about class size seems to me to be valid only if a personnel committee makes the mistake of using ratings to compare teachers rather than as a measure of teaching effectiveness. There is ample evidence that most teachers teach better in small classes. Teachers of small classes require more papers, encourage more discussion, and are more likely to use essay questions on examinations—all of which are likely to contribute to student learning and thinking. Thus, on average, small classes should be rated higher than large classes.[1]

Grading bias, however, is a more serious problem. I have little doubt that giving higher grades can raise ratings if one can convince students that they have learned more than is typical. But students are not so likely to be positively affected if an ineffective teacher seems to be trying to buy good ratings with easy grades. In fact, the attempt may boomerang. A former faculty member whose grades were the highest in my department received the lowest student ratings; Abrami, Dickens, Perry, and Leventhal (1980) presented more systematic evidence of the negative effect of giving undeserved higher grades.

The effect of easy grading may well depend on the institution. Clark and Trow (1966) demonstrated that colleges and universities differ in their dominant cultures: some emphasizing academic values, others emphasizing social and collegiate values. If students have primarily chosen a college to have a good time, easy teachers may be more highly appreciated than in institutions with stronger academic cultures.

Whether or not student ratings are positively affected by grading leniency, the effect on a promotion committee's judgments is likely to be much more negative if the committee perceives the grading pattern to be higher than normal. Even when Sullivan (1974) had convincing evidence that students in his programmed learning class achieved more than students in conventional classes, he encountered fierce hostility from his colleagues about giving higher grades. Faculty members and administrators are concerned about possible grade inflation. Good student ratings accompanied by a higher than normal grade distribution are likely to be a ticket to termination before tenure.[2]

Can Something Be Done to Prevent the Success of Those Who Attempt to Buy Higher Ratings From Students With High Grades?

Greenwald and Gillmore (1997) suggest that only the grading-bias hypothesis can explain four patterns in correlational data and thus justify the use of a statistical correction. Unfortunately, their argument that only the grading-bias hypothesis can account for their four findings seems to me to be flawed. I examine each in turn.

[1] Greenwald and Gillmore (1997) are concerned that even though a teacher who teaches a small class may be more effective, this makes for an unfair advantage when that teacher is compared with a teacher of a large class. I argue that the mistake is in making such comparisons rather than in a bias in the student ratings.

[2] Greenwald (1997) suggests that most personnel committees are not aware of differences in grading standards. It may be that this varies among institutions. Certainly it has come up a number of times in my experience as a department chair and a committee member. I asked one of the senior members of the faculty at the University of Michigan who not only had experience on committees at the University of Michigan but also had chaired a department at another major university about his experience, and he said that grading leniency did come up frequently when discussing a faculty member's teaching.

Positive Grades–Ratings Relationships Within Classes

Here, the assumption is that the teacher is equally effective for all students within a class. In fact, there are numerous studies that have shown attribute–treatment interactions. With specific reference to the within-class correlations, Remmers (1928) suggested, and Elliott (1950) showed, that within-class correlations between grades and student ratings are a function of the level at which the instructor pitches the class. If the instructor teaches primarily to the better students (as many teachers do), then these students achieve more than expected and rate the instructor more highly than do other students, resulting in a positive correlation. By contrast, in a class where the teacher helps poorer students achieve more than predicted, these students give the instructor higher ratings, resulting in a negative correlation between ratings and grades. Greenwald and Gillmore's (1997) results support the common impression that many teachers teach to the better students; in fact, it has not been long since first-year courses (particularly in the sciences) were designed to weed out students who did not belong in those disciplines.

Stronger Grades–Ratings Relationships With Relative, Rather Than Absolute, Measures of Expected Grade

If students feel that they are learning more in a particular class than they are in other classes, it should not be surprising that they will rate teaching effectiveness higher in the former class. Why, then, is not the actual grade expected as highly correlated with ratings as the relative grade? This is likely to be true if teachers strike a chord with some students whose performance in other classes is average or below average. These students will rate the teacher highly and expect their grades to be higher than normal, but the actual expected grades will still not be *A*s. Again, the teaching-effectiveness hypothesis is not disconfirmed by this result.

Grade-Related Halo Effect in Judging Course Characteristics

I admire Greenwald and Gillmore's (1997) ingenuity in thinking of this analysis. Nevertheless, their argument seems to me to be irrelevant to the validity of between-course ratings. As Greenwald and Gillmore point out, students tend to blame the instructor if they fail to learn; thus, it is not surprising that they find fault with many characteristics of the teacher. Those who are having the most difficulty are most likely to blame the situation, resulting in a negative halo. Nevertheless, it may be stretching my attribute–treatment interaction hypothesis too far to explain the halo within, but not between, classes.

Because the focus of the ratings is on overall teaching effectiveness, one should not be surprised to find a halo effect. The appropriate question is as follows: Does the halo effect invalidate students' overall ratings of teaching effectiveness? It probably does to the degree that concern for students' learning and other positive teacher characteristics are overweighted by students in their overall judgment. Thus, those students (frequently the less able) who feel that the teacher does not care about their learning develop a negative halo, whereas those who feel that the teacher cares about them develop a positive halo. However, this does not bear on the validity of the overall rating. In fact, as d'Apollonia and Abrami (1997) point out, the halo effect may increase validity.

Negative Grades–Workload Relationship Between Classes

The negative relationship between grades and workload is not directly relevant to the issue of grading leniency. Part of the relationship is probably due to aggregating data across departments. Science departments tend to give lower grades than humanities and social science departments and are perceived by students as requiring more work (Cashin & Sixbury, 1993; Centra, 1993). Within most departments, there are also ineffective teachers who, feeling alienated from their students, require more work and then blame their students for not meeting the teachers' standards.

Greenwald and Gillmore (1997) assume that hours worked should relate to learning and grades. They probably do. Unfortunately, the relationship is not a simple one. In general, one would expect that students who are having difficulty will spend more time studying than will those who have better background knowledge. This is likely to result in better learning for the less able students but is not likely to result in the kind of positive relationship between workload and grades that Greenwald and Gillmore expected.

Although the workload–grades relationship does not involve student ratings, Greenwald and Gillmore (1997) apparently draw the implication that low-workload courses will be given high ratings. In interviews of students, I have found that often the workload is heavy because the teacher has been ineffective—assignments are unclear, lectures are disorganized, and tests require memorization of definitions and a myriad of specific facts. Thus, Greenwald and Gillmore need to differentiate between hours spent compensating for poor instruction and work that is constructive in promoting learning and increasing motivation. Greenwald and Gillmore could distinguish between these two kinds of "work," I believe, by looking at ratings on such items as "I increased my interest in this field." Again, the teaching-effectiveness hypothesis is not disconfirmed.

Conclusion

Both the grading-bias and teaching-effectiveness hypotheses can account for Greenwald and Gillmore's (1997) findings. Nonetheless, I agree with them that grading leniency can sometimes affect ratings. If the correlation between mean grades and ratings were due only to intentional efforts to get higher ratings, a statistical correction would be appropriate. However, there are at least two kinds of cases in which such a correction would be inappropriate—the excellent teacher whose students' achievement merits higher grades and the poorer teacher whose grades are unjustifiably low. For most teachers, the correction would make little difference. Just as in controlling students' cheating, evaluators should focus on preventive measures rather than implementing measures that will punish effective teachers as well as those who cheat.

Preventing Cheating

What can be done to reduce the sort of desperation that leads to cheating? Clearly, the most desirable measure would be to increase teachers' competence so that student ratings are validly positive, thus reducing the temptation to cheat. This implies strategies such as better preparation for college teaching in graduate school, better orientation and training during the first years of teaching, and collecting student ratings early in the term and discussing them with a consultant or fellow teacher.

The temptation to cheat also may be affected by faculty members' confidence that the judgments will be fair. To make sure that contextual variables influencing ratings are taken into account, personnel committees should consider teachers' own statements about the goals they were trying to achieve, how they went about achieving them, and the contextual conditions that might have influenced success.[3] As Cashin (1995) suggested in his review of the research on correlations between expected grades and ratings, the best method of control is to review graded course materials to judge whether the standards are appropriate.

One would like the committee's judgments to be based on valid evidence. Student ratings are valid, but all of the authors in this *Current Issues* section agree that they should be supplemented with other evidence. Yet, as Marsh and Roche (1997) point out, there is little research on the validity of other sources of evidence. Clearly, such research is needed.

The Validity of Use of Ratings in Personnel Decisions

The authors of the articles in this *Current Issues* section agree that student ratings are the most valid and practical source of data on teaching effectiveness. But, as I noted earlier, these data must then be interpreted by faculty or administrators who must make decisions about promotions and merit pay increases.

I contend that the specific questions used, the use of global versus factor scores, the possible biasing variables, and so forth are relatively minor problems. The major validity problem is in the use of the ratings by personnel committees and administrators (Franklin & Theall, 1989).

No matter how valid the evidence provided by students may be, it is almost certainly more valid than many personnel committees give it credit for being. I have participated in more than 1,000 reviews of faculty members for promotions or merit pay increases. In my opinion, many committees seem to make sensible use of student rating results, but all too often, I have heard student ratings dismissed with such phrases as "He's not a good researcher—obviously he can't be an excellent teacher," "You can't expect students to know which teachers were good until they've been out of college a few years," or "All students want are some jokes and an easy grade." Whatever the reason, student ratings of teaching are often not given heavy weight in promotion decisions.

Although I believe that a statistical adjustment of ratings, such as Greenwald and Gillmore (1997) suggest, may result in lower, rather than higher, validity, it may increase the credibility of the ratings. If it thus contributes to better weighting of ratings in personnel decisions, I'm for it.

Almost as bad as dismissal of student ratings, however, is the opposite problem—attempting to compare teachers with one another by using numerical means or medians. Comparisons of ratings in different classes are dubious not only because of between-classes differences in the students but also because of differences in goals, teaching methods, content, and a myriad of other variables. Moreover, as I suggested earlier, comparisons are not needed for personnel decisions. To the degree that student ratings enter into such decisions, faculty members can be reliably allocated to three or four categories (as d'Apollonia and Abrami [1997] suggest) by simply looking at the distribution of student ratings: How many students rated the teacher as very good or excellent? How many students were dissatisfied?

What Can Be Done to Improve the Validity of the Use of Student Ratings?

Presumably the result educators would like to achieve is appropriate recognition of teaching in personnel decisions, and until those making the decisions become more sophisticated, the nature of the instrument and possible biases are not likely to make significant differences. Research at the University of Michigan on the use of ratings in personnel decisions has used simulated dossiers rated by individual members of committees determining promotions (Lin et al., 1984; Salthouse et al., 1978). But as far as I know, there has been no research on the actual decision-making processes in the committees. It would be difficult, but perhaps not impossible, to obtain permission to carry out observational studies of actual meetings of such committees. If this proves to be impossible, it should be possible to carry out research using simulated meetings in which experimental variations could be tested.

[3] There are probably some classrooms where no one could get top ratings! I once taught a spring class in which the room was unbearably hot if the windows were closed and unbearably noisy from the jackhammers nearby when the windows were open.

If one were to carry out a program of such research with some design that enabled one to discriminate more valid from less valid outcomes, I would not be surprised to find that one would emerge with results similar to those in studies of medical diagnosis and mortality predictions. Either a computer program or a pooled judgment of physicians tends to be superior to predictions of individual physicians. However, the combination of the computer program and the physicians is better yet (Yates, 1994). Thus, I can envision a time when promotion decisions are made by using a weighted combination of Marsh and Roche's (1997) factors along with the pooled judgment of well-trained committee members.

That time is not near, and in the meantime, researchers need to improve the quality of the data presented. The research of Greenwald and Gillmore (1997), Marsh and Roche (1997), d'Apollonia and Abrami (1997), and others has contributed greatly to understanding student ratings of instruction, but in addition to research on student ratings, research is needed on ways of teaching students to be more sophisticated evaluators as well as ways that the experience of filling out the rating form can become more educational for students. For example, qualitative research on what goes on in students' minds when they are filling out evaluations would provide a better idea of whether they are analyzing their own learning or are simply discharging a boring chore. What kinds of items, what kinds of structure for ratings, and what balance of ratings and open-ended questions would stimulate more thought?

Student rating forms have mostly been developed using the approach of "dust bowl empiricism," that is, get a number of items about teaching and see what works. During the 1950s, I tried to collect every student rating form then in use in the United States, and Isaacson et al. (1963) then factor analyzed all of the items I had gathered. I still believe this was a useful approach, but in the 1990s, there are much better theories of cognition and motivation, and student rating forms should now better reflect those theories. Although I have stressed that validity of use is the key issue, researchers should also be looking, as Marsh and Roche (1997) have, at construct validity with respect to theories of teaching.

Even this, however, probably will not be sufficient to handle all the modes of teaching. The increasing use of technology, virtual universities, studio teaching, clinical teaching, cooperative learning, and service learning represents important aspects of education, and we very likely need a variety of forms and items to accommodate such differences.

For summative evaluations by personnel committees, I like the method developed by Hoyt, Owens, Cashin, and others at the Center for Faculty Evaluation and Development at Kansas State University—IDEA (Instructional Development and Effectiveness Assessment). The IDEA form asks students to rate their progress on each of 10 instructional goals—a method that not only provides information that goes beyond the teacher's conformity to a naive stereotype of good teaching but also is educational in broadening students' conceptions of what the aims of education are. Students may not always be able to accurately assess their own progress, but if asked, they do know whether a course mainly required memorization or thinking, and they should know whether a course increased their interest in further learning in that subject-matter area. Use of such items about goals appropriately leaves to the personnel committee the judgment as to which goals are most important for a particular course in the context of the overall objectives of the department and the university. Student ratings of their attainment of educational objectives not only provide better data for personnel committees but also stimulate both students and teachers to think about their objectives—something that is educational in itself.

Researchers also need to study what teachers can do to help students become more sophisticated raters. As I have pointed out, many faculty members and students have rather limited notions of the goals of education and of what is conducive to learning that will last and be used. Thus, faculty need to be educated, and then they need to be encouraged to explain to their students why the requirements they make and the procedures they use are likely to contribute to better learning.

Most of all, research is needed on how to train members of personnel committees to be better evaluators, and research is needed on ways of communicating the results of student evaluations to improve the quality of their use. I was pleased to note that at the 1997 meeting of the American Educational Research Association, some papers were beginning to address these problems of use. For example, Jennifer Franklin (personal communication, March 26, 1997) reported that she is now presenting the ratings and the confidence levels in graphic form to overcome the problem of misinterpretation of numerical means and norms. Katherine Ryan (1997) studied faculty members' views of different reporting approaches and found that faculty members would prefer a standards-based approach rather than norm-referenced reports. As d'Apollonia and Abrami (1997) note, Abrami and I have argued (McKeachie, 1996) that the use of norms not only leads to comparisons that are invalid but also is damaging to the motivation of the 50% of faculty members who find that they are below average. Moreover, presentation of numerical means or medians (often to two decimal places) leads to making decisions based on small numerical differences—differences that are unlikely to distinguish between competent and incompetent teachers.

With respect to my plea for training members of personnel committees, Villaescusa, Franklin, and Aleamoni (1997) reported that a workshop for faculty and administrators improved knowledge and opinions about student ratings. Unfortunately, Ryan (1997) found that most faculty members would not be interested in attending such a workshop. We need research on methods, in addition to workshops, that can help increase the valid use of ratings in personnel decisions. Could such committees be persuaded to accept consultants who would assist them in interpreting student ratings but not take part in the actual decision making?

There now is ample evidence of ways in which teaching can be improved. The problem is how to get research findings into use. Therefore, research and theory is needed not only on the nature and measurement of good teaching but also on the problems of getting theory into use—use in training teachers; use in personnel decisions; and use in methods of collecting data about teaching, such as portfolios, classroom observations, assessments of syllabi, tests, and course materials, and student rating forms.

Conclusion

As Herb Simon (1997) recently said, "Learning is ultimately a human activity, regardless of the technology used." Students will continue to be those most affected by teaching. Therefore, student ratings will continue to be useful.

I end by considering once again Greenwald's (1997) experience that initiated this set of articles. He was surprised that he received markedly lower ratings in one course than in another course that he had taught in the same way. Had I been consulting with him about the ratings, I would have said something like this:

> Tony, classes differ. Effective teaching is not just a matter of finding a method that works well and using it consistently. Rather, teaching is an interactive process between the students and the teacher. Good teaching involves building bridges between what is in your head and what is in the students' heads. What works for one student or for one class may not work for others. Next time, get some ratings early in the term, and if things are not going well, let's talk about varying your strategies.

Fortunately, I was not his consultant, and the result was the series of research studies he and Gillmore (1997) reported as well as the initiation of this group of articles.

REFERENCES

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? *Journal of Educational Psychology, 72,* 107–118.

Aleamoni, L. M. (1978). The usefulness of students' evaluations in improving college teaching. *Instructional Science, 7,* 95–105.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited* (IDEA Paper No. 32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Sixbury, G. R. (1993). *Comparative data by academic field* (IDEA Tech. Rep. No. 8). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco: Jossey-Bass.

Clark, B., & Trow, M. (1966). The organizational context. In T. M. Newcomb & E. K. Wilson (Eds.), *College peer groups: Problems and prospects for research* (pp. 17–70). Chicago: Aldine.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13,* 321–341.

Cross, D. (1958). *An investigation of the relationships between students' expressions of satisfaction with certain aspects of the college classroom situation and their achievement on the final examination.* Unpublished honors thesis, University of Michigan.

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

Elliott, D. H. (1950). Characteristics and relationships of various criteria of colleges and university teaching. *Purdue University Studies in Higher Education, 70,* 5–61.

Franklin, J., & Theall, M. (1989, April). *Who read ratings: Knowledge, attitude and practice of users of student ratings of instruction.* Paper presented at the 70th annual meeting of the American Educational Research Association, San Francisco.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52,* 1182–1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209–1217.

Hoyt, D. P., & Cashin, W. E. (1977). *Development of the IDEA system* (IDEA Tech. Rep. No. 1). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Isaacson, R. L., McKeachie, W. J., & Milholland, J. M. (1963). Correlation of teacher personality variables and student ratings. *Journal of Educational Psychology, 54,* 110–117.

Lin, Y.-G., McKeachie, W. J., & Tucker, D. G. (1984). The use of student ratings in promotion decisions. *Journal of Higher Education, 55,* 583–589.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76,* 707–754.

Marsh, H. W., & Overall, J. U. (1979). Long-term stability of students' evaluations: A note on Feldman's "Consistency and Variability Among College Students in Rating Their Teachers and Courses." *Research in Higher Education, 10,* 139–147.

Marsh, H. W., & Roche, L. A. (1993). The use of student evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30,* 217–251.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52,* 1187–1197.

McKeachie, W. J. (1951). Anxiety in the college classroom. *Journal of Educational Research, 45,* 135–160.

McKeachie, W. J. (1996). Do we need norms of student ratings to evaluate faculty? *Instructional Evaluation and Faculty Development, 14,* 14–17.

McKeachie, W. J., Guetzkow, H., & Kelly, E. L. (1954). An experimental comparison of recitation, discussion and tutorial methods in college teaching. *Journal of Educational Psychology, 45,* 224–232.

McKeachie, W. J., Lin, Y.-G., Daugherty, M., Moffett, M., Neigler, C., Nork, J., Walz, M., & Baldwin, R. (1980). Using student ratings and consultation to improve instruction. *British Journal of Educational Psychology, 50,* 168–174.

McKeachie, W. J., Lin, Y.-G., & Mann, W. (1971). Student ratings of teaching effectiveness: Validity studies. *American Educational Research Journal, 8,* 435–445.

McKeachie, W. J., & Pintrich, P. (1991). Program on classroom teaching and learning strategies. In J. S. Stark & W. J. McKeachie (Eds.), *Final report: National Center for Research to Improve Postsecondary Teaching and Learning* (pp. 41–59). Ann Arbor: University of Michigan, School of Education.

McKeachie, W. J., & Solomon, D. (1958). Student ratings of instructors: A validity study. *Journal of Educational Research, 51,* 379–382.

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75,* 138–149.

Murray, H. G. (1997, March). *Classroom teaching behaviors and student instructional ratings: How do good teachers teach?* McKeachie Award address presented at the 78th annual meeting of the American Educational Research Association, Chicago.

Murray, H. G., & Lan, M. (1997). The relationship between active participation and student learning. *STLHE/SAPES, 20,* 7–10.

Murray, H. G., & Smith, T. A. (1989, April). *Effects of midterm behavioral feedback on end-of-term ratings of instructor effectiveness.* Paper presented at the 70th annual meeting of the American Educational Research Association, San Francisco.

Parsons, T. S. (1957). A comparison of learning by kinescope, correspondence study, and customary classroom procedures. *Journal of Educational Psychology, 48,* 27–40.

Remmers, H. H. (1928). The relationships between students' marks and students' attitudes toward instructors. *School and Society, 28,* 759–760.

Renaud, R. D., & Murray, H. G. (1997, March). *Factorial validity of student ratings of intruction.* Paper presented at the 78th annual meeting of the American Educational Research Association, Chicago.

Ryan, K. E. (1997, March). *Making student ratings comprehensible to faculty: A review of alternative reporting approaches.* Paper presented at the 78th annual meeting of the American Educational Research Association, Chicago.

Salthouse, T. A., McKeachie, W. J., & Lin, Y.-G. (1978). An experimental investigation of factors affecting university promotion decisions. *Journal of Higher Education, 49,* 177–183.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 244–271). Beverly Hills, CA: Sage.

Simon, H. (1997, March). *The future of education in the 21st century.* Paper presented at the Celebration of the 50th Anniversary of the Founding of the American Institutes of Research, Washington, DC.

Sullivan, A. M. (1974). Psychology and teaching. *Canadian Journal of Behavioral Science, 6,* 1–29.

Villaescusa, T., Franklin, J., & Aleamoni, L. (1997, March). *Improving the interpretation and use of student ratings: A training approach.* Paper presented at the 78th annual meeting of the American Educational Research Association, Chicago.

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381–409). New York: Wiley.

Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors
Author(s): Scott E. Carrell and James E. West
Source: *Journal of Political Economy*, Vol. 118, No. 3 (June 2010), pp. 409-432
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/10.1086/653808
Accessed: 31-03-2016 14:03 UTC

Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors

Scott E. Carrell

University of California, Davis and National Bureau of Economic Research

James E. West

U.S. Air Force Academy

In primary and secondary education, measures of teacher quality are often based on contemporaneous student performance on standardized achievement tests. In the postsecondary environment, scores on student evaluations of professors are typically used to measure teaching quality. We possess unique data that allow us to measure relative student performance in mandatory follow-on classes. We compare metrics that capture these three different notions of instructional quality and present evidence that professors who excel at promoting contemporaneous student achievement teach in ways that improve their student evaluations but harm the follow-on achievement of their students in more advanced classes.

Thanks go to U.S. Air Force Academy personnel, R. Schreiner, W. Bremer, R. Fullerton, J. Putnam, D. Stockburger, K. Carson, and P. Egleston, for assistance in obtaining the data for this project and to Deb West for many hours entering data from archives. Thanks also go to F. Hoffmann, H. Hoynes, C. Hoxby, S. Imberman, C. Knittel, L. Lefgren, M. Lovenheim, T. Maghakian, D. Miller, P. Oreopoulos, M. Page, J. Rockoff, and D. Staiger and all seminar participants at the American Education Finance Association meetings, Clemson University, Duke University, NBER Higher Ed Working Group, Stanford University, and University of California, Berkeley and Davis for their helpful comments. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the U.S. Air Force, the Department of Defense, or the U.S. government.

[*Journal of Political Economy*, 2010, vol. 118, no. 3]

> A weak faculty operates a weak program that attracts weak students. (Koerner 1963)

I. Introduction

Conventional wisdom holds that "higher-quality" teachers promote better educational outcomes. Since teacher quality cannot be directly observed, measures have largely been driven by data availability. At the elementary and secondary levels, scores on standardized student achievement tests are the primary measure used and have been linked to teacher bonuses and terminations (Figlio and Kenny 2007). At the postsecondary level, student evaluations of professors are widely used in faculty promotion and tenure decisions. However, teachers can influence these measures in ways that may reduce actual student learning. Teachers can "teach to the test." Professors can inflate grades or reduce academic content to elevate student evaluations. Given this, how well do each of these measures correlate with the desired outcome of actual student learning?

Studies have found mixed evidence regarding the relationship between observable teacher characteristics and student achievement at the elementary and secondary education levels.[1] As an alternative method, teacher "value-added" models have been used to measure the total teacher input (observed and unobserved) to student achievement. Several studies find that a one-standard-deviation increase in teacher quality improves student test scores by roughly one-tenth of a standard deviation (Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Aaronson et al. 2007; Kane, Rockoff, and Staiger 2008). However, recent evidence from Kane and Staiger (2008) and Jacob, Lefgren, and Sims (2010) suggests that these contemporaneous teacher effects may decay relatively quickly over time,[2] and Rothstein (2010) finds evidence that the nonrandom place-

[1] Jacob and Lefgren (2004) find that principal evaluations of teachers were the best predictor of student achievement; Clotfelter, Ladd, and Vigdor (2006, 2007) find evidence that National Board certification and teacher licensure test scores positively predict teacher effectiveness; Dee (2004, 2005) finds that students perform better with same race and gender teachers; and Harris and Sass (2007) find some evidence that teacher professional development is positively correlated with student achievement in middle and high school math. Summers and Wolfe (1977), Cavalluzzo (2004), Vandevoort, Amrein-Beardsley, and Berliner (2004), and Goldhaber and Anthony (2007) find positive effects from teachers certified by the National Board for Professional Teaching Standards. See also Hanushek (1971), Murnane (1975), Summers and Wolfe (1977), Ehrenberg and Brewer (1994), Ferguson and Ladd (1996), Boyd et al. (2006), and Aaronson, Barrow, and Sander (2007).

[2] Jacob et al. (2010) find that 20 percent of the contemporaneous effects persist into the subsequent year. Rothstein (2010) finds that roughly 50 percent persists into year 1 and none persists into year 2 for mathematics courses.

ment of students to teachers may bias value-added estimates of teacher quality.[3]

Even less is known about how the quality of instruction affects student outcomes at the postsecondary level.[4] Standardized achievement tests are not given at the postsecondary level, and grades are not typically a consistent measure of student academic achievement because of heterogeneity of assignments/exams and the mapping of those assessment tools into final grades across individual professors. Additionally, it is difficult to measure how professors affect student achievement because students generally "self-select" their course work and their professors. For example, if better students tend to select better professors, then it is difficult to statistically separate the teacher effects from the selection effects. As a result, the primary tool used by administrators to measure professor teaching quality is scores on subjective student evaluations, which are likely endogenous with respect to (expected) student grades.

To address these various measurement and selection issues in measuring teacher quality, our study uses a unique panel data set from the United States Air Force Academy (USAFA) in which students are randomly assigned to professors over a wide variety of standardized core courses. The random assignment of students to professors, along with a vast amount of data on both professors and students, allows us to examine how professor quality affects student achievement free from the usual problems of self-selection. Furthermore, performance in USAFA core courses is a consistent measure of student achievement because faculty members teaching the same course use an identical syllabus and give the same exams during a common testing period.[5] Finally, USAFA students are required to take and are randomly assigned to numerous follow-on courses in mathematics, humanities, basic sciences, and engineering. Performance in these mandatory follow-on courses is arguably a more persistent measurement of student learning. Thus, a distinct advantage of our data is that even if a student has a particularly poor introductory course professor, he or she still is required to take the follow-on related curriculum.[6]

[3] However, Kane and Staiger (2008) show that controlling for prior year test scores produces unbiased estimates in the presence of self-selection.

[4] Hoffmann and Oreopoulos (2009) find that perceived professor quality, as measured by teaching evaluations, affects the likelihood of a student dropping a course and taking subsequent courses in the same subject. Other recent postsecondary studies have focused on the effectiveness of part-time (adjunct) professors. See Ehrenberg and Zhang (2005) and Bettinger and Long (2006).

[5] Common testing periods are used for freshman- and sophomore-level core courses. All courses are taught without the use of teaching assistants, and faculty members are required to be available for appointments with students from 7:30 a.m. to 4:30 p.m. each day classes are in session.

[6] For example, students of particularly bad Calculus I instructors must still take Calculus II and six engineering courses, even if they decide to be a humanities major.

These properties enable us to measure professor quality free from selection and attrition bias. We start by estimating professor quality using teacher value-added in the contemporaneous course. We then estimate value-added for subsequent classes that require the introductory course as a prerequisite and examine how these two measures covary. That is, we estimate whether high- (low-) value-added professors in the introductory course are high- (low-) value-added professors for student achievement in follow-on related curriculum. Finally, we examine how these two measures of professor value-added (contemporaneous and follow-on achievement) correlate with professor observable attributes and student evaluations of professors. These analyses give us a unique opportunity to compare the relationship between value-added models (currently used to measure primary and secondary teacher quality) and student evaluations (currently used to measure postsecondary teacher quality).

Results show that there are statistically significant and sizable differences in student achievement across introductory course professors in both contemporaneous and follow-on course achievement. However, our results indicate that professors who excel at promoting contemporaneous student achievement, on average, harm the subsequent performance of their students in more advanced classes. Academic rank, teaching experience, and terminal degree status of professors are negatively correlated with contemporaneous value-added but positively correlated with follow-on course value-added. Hence, students of less experienced instructors who do not possess a doctorate perform significantly better in the contemporaneous course but perform worse in the follow-on related curriculum.

Student evaluations are positively correlated with contemporaneous professor value-added and negatively correlated with follow-on student achievement. That is, students appear to reward higher grades in the introductory course but punish professors who increase deep learning (introductory course professor value-added in follow-on courses). Since many U.S. colleges and universities use student evaluations as a measurement of teaching quality for academic promotion and tenure decisions, this latter finding draws into question the value and accuracy of this practice.

These findings have broad implications for how students should be assessed and teacher quality measured. Similar to elementary and secondary school teachers, who often have advance knowledge of assessment content in high-stakes testing systems, all professors teaching a given course at USAFA have an advance copy of the exam before it is given. Hence, educators in both settings must choose how much time to allocate to tasks that have great value for raising current scores but may have little value for lasting knowledge. Using our various measures of quality to rank-order professors leads to profoundly different results. As an illustration, the introductory calculus professor in our sample who ranks dead last in deep learning ranks sixth and seventh best in student evaluations and contemporaneous value-added, respectively. These findings support recent research by Barlevy and Neal (2009), who propose an incentive pay scheme that links teacher compensation to the ranks of their students within appropriately defined comparison sets and requires that new assessments consisting of entirely new questions be given at each testing date. The use of new questions eliminates incentives for teachers to coach students concerning the answers to specific questions on previous assessments.

The remainder of the paper proceeds as follows: Section II reviews the empirical setting. Section III presents the methods and results for professor value-added models. Section IV examines how the observable attributes of professors and student evaluations of instructors are correlated with professor value-added. Section V presents concluding remarks.

II. Empirical Setting

The U.S. Air Force Academy is a fully accredited undergraduate institution of higher education with an approximate enrollment of 4,500 students. There are 32 majors offered including the humanities, social sciences, basic sciences, and engineering. Applicants are selected for admission on the basis of academic, athletic, and leadership potential. All students attending USAFA receive a 100 percent scholarship to cover their tuition, room, and board. Additionally, each student receives a monthly stipend of $845 to cover books, uniforms, computer, and other living expenses. All students are required to graduate within 4 years[7] and serve a 5-year commitment as a commissioned officer in the U.S. Air Force following graduation.

Approximately 40 percent of classroom instructors at USAFA have terminal degrees, as one might find at a university where introductory course work is often taught by graduate student teaching assistants. However, class sizes are very small (average of 20), and student interaction with faculty members is encouraged. In this respect, students' learning experiences at USAFA more closely resemble those of students who attend small liberal arts colleges.

Students at USAFA are high achievers, with average math and verbal Scholastic Aptitude Test (SAT) scores at the 88th and 85th percentiles

[7] Special exceptions are given for religious missions, medical "setbacks," and other instances beyond the control of the individual.

of the nationwide SAT distribution.[8] Students are drawn from each congressional district in the United States by a highly competitive process, ensuring geographic diversity. According to the National Center for Education Statistics (http://nces.ed.gov/globallocator/), 14 percent of applicants were admitted to USAFA in 2007. Approximately 17 percent of the sample is female, 5 percent is black, 7 percent is Hispanic, and 6 percent is Asian. Twenty-six percent are recruited athletes, and 20 percent attended a military preparatory school. Seven percent of students at USAFA have a parent who graduated from a service academy and 17 percent have a parent who previously served in the military.

A. *The Data Set*

Our data set consists of 10,534 students who attended USAFA from the fall of 2000 through the spring of 2007. Student-level pre-USAFA data include whether students were recruited as athletes, whether they attended a military preparatory school, and measures of their academic, athletic, and leadership aptitude. Academic aptitude is measured through SAT verbal and SAT math scores and an academic composite computed by the USAFA admissions office, which is a weighted average of an individual's high school grade point average (GPA), class rank, and the quality of the high school attended. The measure of pre-USAFA athletic aptitude is a score on a fitness test required by all applicants prior to entrance.[9] The measure of pre-USAFA leadership aptitude is a leadership composite computed by the USAFA admissions office, which is a weighted average of high school and community activities (e.g., student council officer, Eagle Scout, captain of a sports team, etc.).

Our primary outcome measure consists of a student-level census of all courses taken and the percentage of points earned in each course. We normalize the percentage of points earned within a course/semester to have a mean of zero and a standard deviation of one. The average percentage of points earned in the course is 78.17, which corresponds to a mean GPA of 2.75.

Students at USAFA are required to take a core set of approximately 30 courses in mathematics, basic sciences, social sciences, humanities, and engineering.[10] Table 1 provides a list of the required math, science, and engineering core courses.

[8] See http://professionals.collegeboard.com/profdownload/sat_percentile_ranks_2008.pdf for SAT score distributions.

[9] Barron, Ewing, and Waddell (2000) found a positive correlation between athletic participation and educational attainment, and Carrell, Fullerton, and West (2009) found a positive correlation between fitness scores and academic achievement.

[10] Over the period of our study there were some changes made to the core curriculum at USAFA.

TABLE 1
Required Math and Science Core Curriculum

Course	Description	Credit Hours
Basic sciences:		
Biology 215	Introductory Biology with Lab	3
Chemistry 141 and 142 or 222	Applications of Chemistry I and II	6
Computer Science 110	Introduction to Computing	3
Mathematics 141	Calculus I	3
Mathematics 142 or 152[a]	Calculus II	3
Mathematics 300, 356, or 377[a]	Introduction to Statistics	3
Physics 110[a]	General Physics I	3
Physics 215[a]	General Physics II	3
Engineering:		
Engineering 100	Introduction to Engineering Systems	3
Engineering 210[a]	Civil Engineering—Air Base Design and Performance	3
Engineering Mechanics 120[a]	Fundamentals of Mechanics	3
Aeronautics 315[a]	Fundamentals of Aeronautics	3
Astronautics 310[a]	Introduction to Astronautics	3
Electrical Engineering 215 or 231[a]	Electrical Signals and Systems	3
Total		45

[a] Denotes that Calculus I is required as a prerequisite to the course.

Individual professor-level data were obtained from USAFA historical archives and the USAFA Center for Education Excellence and were matched to the student achievement data for each course taught by section-semester-year.[11] Professor data include academic rank, gender, education level (master of arts or doctorate), years of teaching experience at USAFA, and scores on subjective student evaluations.

Over the 10-year period of our study we estimate our models using student performance across 2,820 separate course-sections taught by 421 different faculty members. Average class size was 20 students, and approximately 38 sections of each course were taught per year. The average number of classes taught by each professor in our sample is nearly seven. Table 2 provides summary statistics of the data.

[11] Owing to the sensitivity of the data, we were able to obtain the professor observable data only for mathematics, chemistry, and physics. Results for physics and chemistry professors can be found in Carrell and West (2008). Because of the large number of faculty in these departments, a set of demographic characteristics (e.g., female assistant professor, doctorate with 3 years of experience) does not uniquely identify an individual faculty member.

TABLE 2
SUMMARY STATISTICS

	Observations	Mean	Standard Deviation
Student-level variables:			
Total course hours	10,534	16.29	7.99
GPA	10,534	2.75	.80
Percentage of points earned in courses (mean)	10,534	78.17	8.45
SAT verbal	10,534	632.30	66.27
SAT math	10,534	663.51	62.80
Academic composite	10,533	12.82	2.13
Leadership composite	10,508	17.30	1.85
Fitness score	10,526	4.66	.99
Female	10,534	.17	.38
Black	10,534	.05	.22
Hispanic	10,534	.07	.25
Asian	10,534	.06	.23
Recruited athlete	10,534	.25	.44
Attended preparatory school	10,534	.20	.40
Professor-level variables:[a]			
Instructor is a lecturer	91	.58	.50
Instructor is an assistant professor	91	.26	.44
Instructor is an associate or full professor	91	.15	.36
Instructor has a terminal degree	91	.31	.46
Instructor's teaching experience	91	3.66	4.42
Number of sections taught	421	6.64	5.22
Class-level variables:[b]			
Class size	2,820	20.28	3.48
Number of sections per course per year	2,820	38.37	12.80
Average class SAT verbal	2,820	631.83	22.05
Average class SAT math	2,820	661.61	27.77
Average class academic composite	2,820	12.84	.73
Student evaluation of professors by section:[c]			
Instructor's ability to provide clear, well-organized instruction was	237	4.48	.70
Value of questions and problems raised by instructor was	237	4.50	.57
Instructor's knowledge of course material was	237	5.02	.58
The course as a whole was	237	4.08	.61
Amount you learned in the course was	237	4.09	.58
The instructor's effectiveness in facilitating my learning in the course was	237	4.42	.69

[a] Observable attribute data are available only for calculus professors.

[b] Class-level data include introductory calculus and follow-on related core courses.

[c] Student evaluation data are for introductory calculus professors only. The number of observations is the number of sections.

B. *Student Placement into Courses and Sections*

Prior to the start of the freshman academic year, students take course placement exams in mathematics, chemistry, and select foreign languages. Scores on these exams are used to place students into the appropriate starting core courses (i.e., remedial math, Calculus I, Calculus II, etc.). Conditional on course placement, the USAFA registrar employs a stratified random assignment algorithm to place students into sections within each course/semester. The algorithm first assigns all female students evenly throughout all offered sections, then places male-recruited athletes, and then assigns all remaining students. Within each group (i.e., female, male athlete, and all remaining males), assignments are random with respect to academic ability and professor.[12] Thus, students throughout their 4 years of study have no ability to choose their professors in required core courses. Faculty members teaching the same course use an identical syllabus and give the same exams during a common testing period. These institutional characteristics assure that there is no self-selection of students into (or out of) courses or toward certain professors.

Although the placement algorithm used by the USAFA registrar should create sections that are a random sample of the course population with respect to academic ability, we employed resampling techniques as in Lehmann and Romano (2005) and Good (2006) to empirically test this assumption. For each section of each core course/semester we randomly drew 10,000 sections of equal size from the relevant introductory course enrollment without replacement. Using these randomly sampled sections, we computed the sums of both the academic composite score and the SAT math score.[13] We then computed empirical *p*-values for each section, representing the proportion of simulated sections with values less than that of the observed section.

Under random assignment, any unique *p*-value is equally likely to be observed; hence the expected distribution of the empirical *p*-values is uniform. We tested the uniformity of the distributions of empirical *p*-values by semester by course using both a Kolmogorov-Smirnov one-sample equality of distribution test and a χ^2 goodness of fit test.[14] As

[12] In-season intercollegiate athletes are not placed into the late-afternoon section, which starts after 3:00 p.m.

[13] We performed resampling analysis on the USAFA classes of 2000–2009. We also conducted the resampling analysis for SAT verbal and math placement scores and found qualitatively similar results. For brevity we do not present these results in the text.

[14] The Kolmogorov-Smirnov test equals $\sup_x |F_n(x) - F(x)|$, where $F_n(x)$ is the empirical cumulative distribution function and $F(x)$ is the theoretical cumulative distribution function;

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \eta_i)^2}{\eta_i},$$

where n_i is the observed frequency in bin i and η_i is the expected frequency in bin i.

TABLE 3
Randomness Checks

	Calculus I		Calculus II	
Professor Characteristic	Academic Composite (1)	SAT Math (2)	Academic Composite (1)	SAT Math (2)
Associate/full professor	.029 (.060)	.002 (.073)	.009 (.060)	−.024 (.056)
Experience	.000 (.005)	.000 (.006)	.002 (.007)	−.003 (.004)
Terminal degree	.031 (.039)	.056 (.040)	.038 (.042)	−.006 (.037)
Empirical *p*-values (mean and standard deviation)	.512 (.311)	.514 (.334)	.503 (.302)	.503 (.315)
Kolmogorov-Smirnov test (no. failed/total tests)	0/20	1/20	0/20	0/20
χ^2 goodness of fit test (no. failed/total tests)	1/20	2/20	0/20	0/20

Note.—Each cell represents regression results in which the dependent variable is the empirical *p*-value from resampling as described in Sec. II.B and the independent variable is the professor characteristic. Because of the collinearity of the regressors, each column represents results for three separate regressions for associate/full professor, experience, and terminal degree. All specifications include semester by year fixed effects. Standard errors are clustered by professor. The empirical *p*-value of each section represents the proportion of the 10,000 simulated sections with values less than that of the observed section. The Kolmogorov-Smirnov and χ^2 goodness of fit test results indicate the number of tests of the uniformity of the distribution of *p*-values that failed at the 5 percent level.

reported in table 3, we rejected the null hypothesis of random placement for only one of 80 course/semester test statistics at the .05 level using the Kolmogorov-Smirnov test and three of 80 course/semester test statistics using the χ^2 goodness of fit test. As such, we found virtually no evidence of nonrandom placement of students into sections by academic ability.

Next, we tested for the random placement of professors with respect to student ability by regressing the empirical *p*-values from resampling by section on professor academic rank, years of experience, and terminal degree status. Results for this analysis are shown in table 3 and indicate that there is virtually no evidence of nonrandom placement of professors into course sections. Of the 36 estimated coefficients, none are statistically significant at the 5 percent level.

Results from the preceding analyses indicate that the algorithm that places students into sections within a course and semester appears to be random with respect to both student and professor characteristics.

C. Are Student Scores a Consistent Measure of Student Achievement?

The integrity of our results depends on the percentage of points earned in core courses being a consistent measure of relative achievement across students. The manner in which student scores are determined at USAFA, particularly in the Math Department, allows us to rule out potential mechanisms for our results. Math professors grade only a small proportion of their own students' exams, vastly reducing the ability of "easy" or "hard" grading professors to affect their students' scores. All math exams are jointly graded by all professors teaching the course during that semester in "grading parties," where Professor A grades question 1 and Professor B grades question 2 for all students taking the course. These aspects of grading allow us to rule out the possibility that professors have varying grading standards for equal student performance. Hence, our results are likely driven by the manner in which the course is taught by each professor.

In some core courses at USAFA, 5–10 percent of the overall course grade is earned by professor/section-specific quizzes and/or class participation. However, for the period of our study, the introductory calculus course at USAFA did not allow for any professor-specific assignments or quizzes. Thus, potential "bleeding heart" professors had no discretion to boost grades or to keep their students from failing their courses. For this reason, we present results in this study for the introductory calculus course and follow-on courses that require introductory calculus as a prerequisite.[15]

III. Professor Value-Added

A. *Empirical Model*

The professor value-added model estimates the total variance in professor inputs (observed and unobserved) in student academic achievement by utilizing the panel structure of our data, where different professors teach multiple sections of the same course across years. We estimate professor value-added using a random effects model. Random effects estimators are minimum variance and efficient but are not typically used in the teacher quality literature because of the stringent requirement for consistency—that teacher value-added be uncorrelated with all other explanatory variables in the model.[16] This requirement is almost certainly violated when students self-select into course work or sections of a given course, but the requirement is satisfied in our context (Raudenbush and Bryk 2002; McCaffrey et al. 2004).

Consider a set of students indexed by $i = 1, \dots, N$ who are randomly placed into sections $s^1 \in \mathbb{S}$ of the introductory course, where the su-

[15] We find qualitatively similar results for chemistry and physics professors in Carrell and West (2008), where the identification is less clean. Chemistry and physics professors were allowed to have section-specific assignments and grade their own students' exams. These results are available on request.

[16] We run a Hausman specification test and fail to reject the null hypothesis that the fixed effects and random effects estimates are equivalent.

perscript 1 denotes an introductory course section. A member of the set of introductory course professors, indexed by $j^1 = 1, \dots, J$, is assigned to each section s^1. In subsequent semesters, each student i is randomly placed into follow-on course sections $s^2 \in \mathbb{S}$, where the superscript 2 denotes a follow-on course section. A member of the set of follow-on course professors, indexed by $j^2 = 1, \dots, J$ (overlapping the set of introductory course professors), is assigned to each section s^2.

The outcomes of student i are given by the following two-equation model:

$$\begin{bmatrix} Y^1_{itj1j2s1s2} \\ Y^2_{itj1j2s1s2} \end{bmatrix} = \begin{bmatrix} X_{its1} & 0 \\ 0 & X_{its2} \end{bmatrix} \begin{bmatrix} \beta^1 \\ \beta^2 \end{bmatrix} + \begin{bmatrix} \gamma_t^{\,1} \\ \gamma_t^{\,2} \end{bmatrix} + \begin{bmatrix} \lambda^1_{j1} + \lambda^1_{j2} + \xi^1_{ts1} + \xi^1_{ts2} \\ \lambda^2_{j2} + \lambda^2_{j1} + \xi^2_{ts2} + \xi^2_{ts1} \end{bmatrix}$$

$$+ \begin{bmatrix} \epsilon^1_{itj1j2s1s2} \\ \epsilon^2_{itj1j2s1s2} \end{bmatrix}, \tag{1}$$

where $Y^1_{itj^1j^2s^1s^2}$ and $Y^2_{itj^1j^2s^1s^2}$ are the normalized percentages of points earned by student i in semester-year t with introductory professor j^1 in section s^1 and follow-on professor j^2 in section s^2. Superscript 1 denotes introductory course achievement and superscript 2 denotes follow-on course achievement. The terms X_{its^1} and X_{its^2} are vectors of student-specific and classroom mean peer characteristics, including SAT math, SAT verbal, academic composite, fitness score, leadership composite, race/ethnicity, gender, recruited athlete, and whether they attended a military preparatory school relevant to sections s^1 and s^2, respectively, in time t. We control for unobserved mean differences in academic achievement or grading standards across time by including course by semester intercepts, $\gamma_t^{\,1}$ and $\gamma_t^{\,2}$.

The λ's are the parameters of primary interest in our study, which measure professor value-added. Specifically, $\lambda^1_{j^1}$ measures the introductory course professor j^1's value-added in the contemporaneous introductory course and $\lambda^2_{j^1}$ measures the introductory course professor j^1's value-added in mandatory follow-on related courses (deep learning). Likewise, $\lambda^2_{j^2}$ measures the follow-on course professor j^2's value-added in the contemporaneous follow-on course and $\lambda^1_{j^2}$ measures the follow-on course professor j^2's value-added in the introductory course. The presence of $\lambda^1_{j^2}$ allows for a second test of random assignment since we expect this effect to be zero. High values of λ indicate that the professor's students perform better on average, and low values of λ indicate lower average achievement. The variance of λ across professors measures the dispersion of professor quality, whether it be observed or unobserved (Rivkin et al. 2005).

The ξ terms are section-specific random effects measuring classroom-level common shocks that are independent across professors j and time t. Specifically, $\xi^1_{ts^1}$ measures the introductory course section-specific shock in the contemporaneous introductory course, and $\xi^2_{ts^1}$ measures the introductory course section-specific common shock in the follow-on course. Likewise, $\xi^2_{ts^2}$ measures the follow-on course section-specific shock in the contemporaneous follow-on course, and $\xi^1_{ts^2}$ measures the follow-on course section-specific common shock in the introductory course. Again, we expect this latter effect to be zero given the random assignment of students to follow-on course sections.

The terms $\epsilon^1_{itj^1j^2s^1s^2}$ and $\epsilon^2_{itj^1j^2s^1s^2}$ are the student-specific stochastic error terms in the introductory and follow-on course, respectively.[17]

B. Results for Introductory Professors

Table 4 presents the full set of estimates of the variances and covariances of the λ's, ξ's, and ϵ's for introductory calculus professors. Covariance elements in the matrix with a value of 0 were set to zero in the model specification.[18]

The estimated variance in introductory professor quality in the contemporaneous introductory course, Var $(\lambda^1_{j^1})$ in row 1, column 1, is 0.0028 (standard deviation [SD] = 0.052) and is statistically significant at the .05 level. This result indicates that a one-standard-deviation change in professor quality results in a 0.05-standard-deviation change in student achievement. In terms of scores, this effect translates into about 0.6 percent of the final percentage of points earned in the course. The magnitude of the effect is slightly smaller but qualitatively similar to those found in elementary school teacher quality estimates (Kane et al. 2008).

When evaluating achievement in the contemporaneous course being taught, the major threat to identification is that the professor value-added model could be identifying a common treatment effect rather than measuring the true quality of instruction. For example, if Professor A "teaches to the test," his students may perform better on exams and earn higher grades in the course, but they may not have learned any more actual knowledge relative to Professor B, who does not teach to the test. In the aforementioned scenario, the contemporaneous model would identify Professor A as a higher-quality teacher than Professor B.

[17] Owing to the complexity of the nesting structure of professors within courses and course sections within professors, we estimate all the above parameters in two separate random effects regression models using Stata's xtmixed command—one model for introductory course professors and another for follow-on course professors.

[18] A unique aspect of our data is that we observe the same professors teaching multiple sections of the same course in each year. In results unreported but available on request, we tested the stability of professor value-added across years and found insignificant variation in the within-professor teacher value-added across years. These results indicate that the existing practice in the teacher quality literature of relying on only year-to-year variation appears to be justified in our setting.

TABLE 4

Variance-Covariance of Professor Value-Added and Course Sections in Contemporaneous and Follow-on Courses

	λ_{j1}^{1} (1)	λ_{j1}^{2} (2)	λ_{j2}^{1} (3)	λ_{j2}^{2} (4)	ξ_{ts1}^{1} (5)	ξ_{ts1}^{2} (6)	ξ_{ts2}^{1} (7)	ξ_{ts2}^{2} (8)	ϵ (9)
1. λ_{j1}^{1}	.0028 (.0001, .0538)								
2. λ_{j1}^{2}	−.0004 (−.0051, .0043)	.0025 (.0006, .0115)							
3. λ_{j2}^{1}	0	0	.0000 (.0000, .0000)						
4. λ_{j2}^{2}	0	0	0	.0186 (.0141, .0245)					
5. ξ_{ts1}^{1}	0	0	0	0	.0255 (.0159, .0408)				
6. ξ_{ts1}^{2}	0	0	0	0	.0251 (.0179, .0324)	.0248 (.0196, .0315)			
7. ξ_{ts2}^{1}	0	0	0	0	0	0	.0000 (.0000, .0000)		
8. ξ_{ts2}^{2}	0	0	0	0	0	0	0	.0100 (.0067, .0149)	
9. ϵ	0	0	0	0	0	0	0	0	.7012 (.6908, .7117)

Note.—The table shows random effects estimates of the variances and covariances for professors, course sections, and students. The model specification includes course by semester fixed effects as well as classroom-level attributes for SAT math, SAT verbal, and academic composite. Individual-level controls include black, Hispanic, Asian, female, recruited athlete, attended a preparatory school, freshman, SAT verbal, SAT math, academic composite, leadership composite, and fitness score and their interactions with a follow-on course indicator. Ninety-five percent confidence intervals are shown in parentheses. The term λ is the random effect of the intro (subscript j^1) or follow-on (subscript j^2) professor in the intro (superscript 1) or follow-on (superscript 2) course; ξ is the random effect of the intro (subscript s^1) or follow-on (subscript s^2) section in the intro (superscript 1) or follow-on (superscript 2) course; and ϵ is the course achievement level error term.

The USAFA's comprehensive core curriculum provides a unique opportunity to test how introductory course professors affect follow-on course achievement free from selection bias. The estimate of Var (λ_{j1}^{2}) is shown in row 2, column 2 of table 4 and indicates that introductory course professors significantly affect follow-on course achievement.[19] The variance in follow-on course value-added is estimated to be 0.0025 (SD = 0.050). The magnitude of this effect is roughly equivalent to that estimated in the contemporaneous course and indicates that a one-standard-deviation change in introductory professor quality results in a 0.05-standard-deviation change in follow-on course achievement.

The preceding estimates of Var (λ_{j1}^{1}) and of Var (λ_{j1}^{2}) indicate that introductory course calculus professors significantly affect student achievement in both the contemporaneous introductory course being taught and follow-on courses. The estimated covariance, Cov $(\lambda_{j1}^{1}, \lambda_{j1}^{2})$, of these professor effects is negative (-0.0004) and statistically insignificant as shown in column 1, row 2 of table 4. This result indicates that being a high- (low-) value-added professor for contemporaneous student achievement is negatively correlated with being a high- (low-) value-added professor for follow-on course achievement. To get a better understanding of this striking result, we next decompose the covariance estimate.

We note that there are two ways in which the introductory professor (i.e., introductory calculus professor) can affect follow-on course achievement (i.e., aeronautical engineering). First, the initial course professor effect can persist into the follow-on course, which we will specify as $\rho\lambda_{j1}^{1}$. Second, the initial course professor can produce value-added not reflected in the initial course, which we will specify as ϕ_{j1}^{2}. One example of ϕ_{j}^{2} would be "deep learning" or understanding of mathematical concepts that are not measured on the calculus exam but would increase achievement in more advanced mathematics and engineering courses. Hence, we can specify λ_{j1}^{2} and its estimated covariance with λ_{j1}^{1} as follows:[20]

$$\lambda_{j1}^{2} = \rho\lambda_{j1}^{1} + \phi_{j1}^{2}, \tag{2}$$

$$\begin{aligned}\mathbb{E}[\lambda_{j1}^{1}\lambda_{j1}^{2}] &= \mathbb{E}[(\lambda_{j1}^{1})(\rho\lambda_{j1}^{1} + \phi_{j1}^{2})] \\ &= \rho \operatorname{Var}(\lambda_{j1}^{1}). \end{aligned} \tag{3}$$

Therefore, Cov $(\lambda_{j1}^{1}, \lambda_{j1}^{2})/$ Var (λ_{j1}^{1}) is a consistent estimate of ρ, the pro-

[19] We estimate λ_{j}^{2} using all the follow-on required courses that require Calculus I as a prerequisite. These courses are listed in table 1.

[20] If ϕ_{j1}^{2} represents value-added from the initial course professor in the follow-on course not reflected in initial course achievement, Cov $(\lambda_{j}^{1}, \phi_{j1}^{2}) = 0$ by construction.

portion of contemporaneous value-added that persists into follow-on course achievement.

Using results from table 4, we estimate ρ at -0.14.[21] Taken jointly, our estimates of Var (λ_{j1}^{1}), Var (λ_{j1}^{2}), and ρ indicate that one set of calculus professors produce students who perform relatively better in calculus and another set of calculus professors produce students who perform well in follow-on related courses, and these sets of professors are not the same.

In figure 1 we show our findings graphically. Figure 1*A* plots classroom average residuals of adjacent sections by professor for introductory and follow-on course achievement as in Kane et al. (2008).[22] Figure 1*B* plots Bayesian shrinkage estimates of the estimated contemporaneous course and follow-on course professor random effects.[23] These results show that introductory course professor value-added in the contemporaneous course is negatively correlated with value-added in follow-on courses (deep learning). On the whole, these results offer an interesting puzzle and, at a minimum, suggest that using contemporaneous student achievement to estimate professor quality may not measure the "true" professor input into the education production function.

C. *Results for Follow-on Course Professors*

Although the primary focus of our study is to examine how introductory professors affect student achievement, our unique data also allow us to measure how follow-on course professors (e.g., Calculus II professors) affect student achievement in both the contemporaneous course (e.g., Calculus II) and the introductory course (e.g., Calculus I), which should

[21] We cannot directly estimate a standard error for ρ within the random effects framework. Since the denominator, Var (λ_{j1}^{1}), must be positive, the numerator, ρ Var (λ_{j1}^{1}), determines the sign of the quotient. As our estimate of ρ Var (λ_{j1}^{1}) is not significantly different from zero, this result is presumably driven by the magnitude of ρ. Using the two-stage least squares methodology by Jacob et al. (2010) to directly estimate ρ and its standard error, we find it to be negative and statistically insignificant.

[22] Classroom average performance residuals are calculated by taking the mean residual when regressing the normalized score in the course by student on course by semester fixed effects, classroom-level attributes for SAT math, SAT verbal, and academic composite; individual-level controls include black, Hispanic, Asian, female, recruited athlete, attended a preparatory school, freshman, SAT verbal, SAT math, academic composite, leadership composite, and fitness score. In results not shown, we estimate our models using a fixed effect framework as in Kane et al. (2008) and Hoffmann and Oreopoulos (2009) and find qualitatively similar results. To isolate professor value-added from section-specific common shocks in the fixed effect framework, we estimate Var (λ_{j1}^{1}) and Var (λ_{j1}^{2}) using pairwise covariances in professor classroom average performance residuals.

[23] The Bayesian shrinkage estimates are a best linear unbiased predictor of each professor's random effect, which take into account the variance (signal to noise) and the number of observations for each professor. Specifically, estimates with a higher variance and a smaller number of observations are shrunk toward zero. See Rabe-Hesketh and Skrondal (2008) for further details.

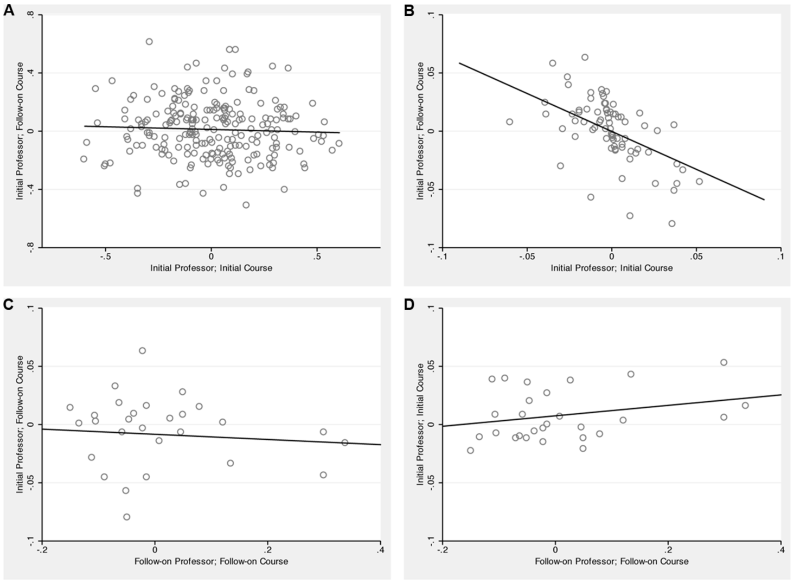

Fig. 1.—Plots of professor effects. *A*, Performance residuals of introductory professor effect in initial course versus introductory professor effect on follow-on course. Classroom average performance residuals were calculated by taking the mean residual when regressing the normalized score in the course by student on course by semester fixed effects, classroom-level attributes for SAT math, SAT verbal, and academic composite; individual-level controls include black, Hispanic, Asian, female, recruited athlete, attended a preparatory school, freshman, SAT verbal, SAT math, academic composite, leadership composite, and fitness score. *B*, Bayesian shrinkage estimates of introductory professor effect in initial course versus introductory professor effect on follow-on course. *C*, Bayesian shrinkage estimates of introductory professor effect of follow-on course versus follow-on professor effect in follow-on course. *D*, Bayesian shrinkage estimates of introductory professor effect in initial course versus follow-on professor effect in follow-on course.

be zero. These results are interesting for two reasons. First, they help test the statistical assumptions of the value-added model as described by Rothstein (2010). Second, we observe a subset of professors in our sample teaching both the introductory and follow-on courses (Calculus I and II). Thus, we are able to examine the correlation between introductory course professor value-added and follow-on course professor value-added.

Rothstein (2010) shows that the assumptions of value-added models are often violated because of the self-selection of students to classrooms and teachers. To illustrate his point, Rothstein (2010) finds that value-added models yield large "effects" of fifth grade teachers on fourth grade test scores. We report estimates for Var $(\lambda^1_{j^2})$, the follow-on professor effect on the initial course grade, in row 3, column 3 of table 4. Consistent with random assignment, we find no evidence that follow-on professors affect introductory course achievement. The estimated variance in the professor random effect is near zero (SD = 0.000002). However, we do find that follow-on professors significantly affect contemporaneous follow-on student achievement. As shown in row 4, column 4, the estimate of Var $(\lambda_{j^2}^2)$ is 0.0185 (SD = 0.136).

To examine the correlation between introductory course professor value-added and follow-on course professor value-added, we show plots of the Bayesian shrinkage estimates in figure 1 for the subset of professors we observe teaching both the introductory and follow-on courses. Figure 1*C* plots fitted values of $\lambda_{j^1}^2$ versus $\lambda_{j^2}^2$ (introductory professor effect on the follow-on course vs. the follow-on professor effect in the follow-on course), and figure 1*D* plots $\lambda_{j^1}^1$ versus $\lambda_{j^2}^2$ (introductory professor effect in the initial course vs. the follow-on professor effect in the follow-on course). These plots yield two interesting findings. First, the clear positive relationship shown in figure 1*D* indicates that professors who are measured as high value-added when teaching the introductory course are also measured as high value-added when teaching the follow-on course. However, the slightly negative and noisy relationship in figure 1*D* indicates that of professors who teach both introductory and follow-on courses, the value-added to the follow-on course produced during the introductory course (deep learning) is uncorrelated with contemporaneously produced value-added in the follow-on course. That is, there appears to be a clear set of professors whose students perform well on sequences of contemporaneous course work, but this higher achievement has little to do with persistent measurable long-term learning.

D. Results for Section-Specific Common Shocks

In both the introductory and follow-on courses, we find significant contemporaneous section-specific common shocks. Although the section-specific common shocks serve primarily to control for section-level variation lest it inappropriately be attributed to professor value-added, the magnitudes and signs of the cross-product common shocks provide a useful check of internal consistency. As expected, the common shock from the introductory course persists into the follow-on course, Var $(\xi_{ts^1}^2) > 0$. In contrast to Rothstein (2010), the common shock in the follow-on course has no effect on introductory course performance, Var $(\xi_{ts^2}^1) = 0$. This is further evidence in support of random student assignment into sections with respect to academic ability.

TABLE 5
Professor Observable Characteristics and Student Evaluations of Professors

	λ^1_{j1} (1)	λ^2_{j1} (2)
	A. Professor Observable Attributes	
Associate/full professor	−.69*	.70*
	(.41)	(.40)
Terminal degree	−.28	.38
	(.27)	(.27)
Greater than 3 years' teaching experience	−.79***	.66**
	(.29)	(.29)
	B. Student Evaluation Scores	
Instructor's ability to provide clear, well-organized instruction was	.51***	−.46**
	(.19)	(.20)
Value of questions and problems raised by instructor was	.70***	−.59**
	(.24)	(.25)
Instructor's knowledge of course material was	.56**	−.44*
	(.24)	(.24)
The course as a whole was	.49**	−.39*
	(.23)	(.23)
Amount you learned in the course was	.59**	−.47*
	(.23)	(.24)
The instructor's effectiveness in facilitating my learning in the course was	.54***	−.45**
	(.20)	(.20)

Note.—Each row by column represents a separate regression in which the dependent variable is the Bayesian shrinkage estimates of the corresponding professor random effects estimated in eq. (1). In all specifications the Bayesian shrinkage estimates were scaled to have a mean of zero and a variance of one. Panel A shows results for modal rank and mean years of teaching experience. Panel B shows results for sample career averages on student evaluations.

* Significant at the .10 level.

** Significant at the .05 level.

*** Significant at the .01 level.

IV. Observable Professor Characteristics and Student Evaluations of Professors

A. *Observable Professor Characteristics*

One disadvantage of the professor value-added model is that it is unable to measure which observable professor characteristics actually predict student achievement. That is, the model provides little or no information to administrators wishing to improve future hiring practices. To measure whether observable professor characteristics are correlated with professor value-added, we regress normalized Bayesian shrinkage estimates from the contemporaneous course, λ^1_{j1}, and follow-on course, λ^2_{j1}, on professor observable attributes.[24] Results are presented in table 5, panel A.

[24] For the professor observable attributes we use mean experience and modal rank. We combine the ranks of associate and full professor, as do Hoffmann and Oreopoulos (2009), because of the small numbers of full professors in our sample. Lecturers at USAFA are typically younger military officers (captains and majors) with master's degrees.

The overall pattern of the results shows that students of less experienced and less qualified professors perform significantly better in the contemporaneous course being taught. In contrast, the students of more experienced and more highly qualified introductory professors perform significantly better in the follow-on courses. Here, we have normalized the shrinkage estimates of professor value-added to have a mean of zero and a standard deviation of one. Thus, in column 1, panel A, the negative coefficient for the associate/full professor dummy variable (-0.69) indicates that shrinkage estimates of contemporaneous value-added among professors are, on average, 0.69 standard deviations lower for senior ranking professors than for lecturers. Conversely, the positive and significant result (0.70) for the associate/full professor dummy variable in column 2 indicates that these same professors teach in ways that enhance student performance in follow-on courses. We find a similar pattern of results for the terminal degree and experience variables.

The manner in which student scores are determined at the USAFA as described in Section II.C allows us to rule out the possibility that higher-ranking professors have higher grading standards for equal student performance. Hence, the preceding results are likely driven by the manner in which the course is taught by each professor.[25]

B. *Student Evaluations of Professors*

Next, we examine the relationship between student evaluations of professors and student academic achievement as in Weinberg, Hashimoto, and Fleisher (2009). This analysis gives us a unique opportunity to compare the relationship between value-added models (currently used to measure primary and secondary teacher quality) and student evaluations (currently used to measure postsecondary teacher quality).

To measure whether student evaluations are correlated with professor value-added, we regress the normalized Bayesian shrinkage estimates from the contemporaneous course, $\lambda^1_{j^1}$, and follow-on course, $\lambda^2_{j^1}$, on career averages from various questions on the student evaluations.[26] Results presented in table 5, panel B, show that student evaluation scores are positively correlated with contemporaneous course value-added but negatively correlated with deep learning.[27] In column 1, results for contemporaneous value-added are positive and statistically significant at the

[25] To test for possible attrition bias in our estimates, we examined whether observable teacher characteristics in the introductory courses were correlated with the probability a student drops out after the first year and whether the student ultimately graduates. Results had various signs, were small in magnitude, and were statistically insignificant.

[26] Again, for ease of interpretation we normalized the Bayesian shrinkage estimates to have a mean of zero and a variance of one.

[27] For brevity, we present results for only a subset of questions; however, results were qualitatively similar across all questions on the student evaluation form.

.05 level for scores on all six student evaluation questions. In contrast, results in column 2 for follow-on course value-added show that all six coefficients are negative, with three significant at the .05 level and three significant at the .10 level

Since proposals for teacher merit pay are often based on contemporaneous teacher value-added, we examine rank orders between our professor value-added estimates and student evaluation scores. We compute rank orders of career average student evaluation data for the question, “The instructor’s effectiveness in facilitating my learning in the course was,” by professor, $r(\omega^1_{j1}) = r_{\omega^1_1}$, and rank orders of the Bayesian shrinkage estimates of introductory professor value-added in the introductory course, $r(\lambda^1_{j1}) = r_{\lambda^1_1}$, and introductory course professor value-added in the follow-on course, $r(\lambda^2_{j1}) = r_{\lambda^2_1}$. Consistent with our previous findings, the correlation between introductory calculus professor value-added in the introductory and follow-on courses is negative, $\mathrm{Cor}(r_{\lambda^1_1}, r_{\lambda^2_1}) = -0.68$. Students appear to reward contemporaneous course value-added, $\mathrm{Cor}(r_{\lambda^1_1}, r_{\omega^1_1}) = 0.36$, but punish deep learning, $\mathrm{Cor}(r_{\lambda^2_1}, r_{\omega^1_1}) = -0.31$. As an illustration, the calculus professor in our sample who ranks dead last in deep learning ranks sixth and seventh best in student evaluations and contemporaneous value-added, respectively.

V. Conclusion

Our findings show that introductory calculus professors significantly affect student achievement in both the contemporaneous course being taught and the follow-on related curriculum. However, these methodologies yield very different conclusions regarding which professors are measured as high quality, depending on the outcome of interest used. We find that less experienced and less qualified professors produce students who perform significantly better in the contemporaneous course being taught, whereas more experienced and highly qualified professors produce students who perform better in the follow-on related curriculum.

Owing to the complexities of the education production function, where both students and faculty engage in optimizing behavior, we can only speculate as to the mechanism by which these effects may operate. Similar to elementary and secondary school teachers, who often have advance knowledge of assessment content in high-stakes testing systems, all professors teaching a given course at USAFA have an advance copy of the exam before it is given. Hence, educators in both settings must choose how much time to allocate to tasks that have great value for raising current scores but may have little value for lasting knowledge.

One potential explanation for our results is that the less experienced professors may adhere more strictly to the regimented curriculum being tested, whereas the more experienced professors broaden the curriculum and produce students with a deeper understanding of the material. This deeper understanding results in better achievement in the follow-on courses. Another potential mechanism is that students may learn (good or bad) study habits depending on the manner in which their introductory course is taught. For example, introductory professors who "teach to the test" may induce students to exert less study effort in follow-on related courses. This may occur because of a false signal of one's own ability or an erroneous expectation of how follow-on courses will be taught by other professors. A final, more cynical, explanation could also relate to student effort. Students of low-value-added professors in the introductory course may increase effort in follow-on courses to help "erase" their lower than expected grade in the introductory course.

Regardless of how these effects may operate, our results show that student evaluations reward professors who increase achievement in the contemporaneous course being taught, not those who increase deep learning. Using our various measures of teacher quality to rank-order teachers leads to profoundly different results. Since many U.S. colleges and universities use student evaluations as a measurement of teaching quality for academic promotion and tenure decisions, this finding draws into question the value and accuracy of this practice.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *J. Labor Econ.* 25 (1): 95–135.

Barlevy, Gadi, and Derek Neal. 2009. "Pay for Percentile." Working Paper no. 2009-09, Fed. Reserve Bank Chicago. http://www.chicagofed.org/webpages/publications/working_papers/2009/wp_09.cfm.

Barron, John M., Bradley T. Ewing, and Glen R. Waddell. 2000. "The Effects of High School Participation on Education and Labor Market Outcomes." *Rev. Econ. and Statis.* 82 (3): 409–21.

Bettinger, Eric, and Bridget Terry Long. 2006. "The Increasing Use of Adjunct Instructors at Public Institutions: Are We Hurting Students?" In *What's Happening to Public Higher Education?* edited by Ronald G. Ehrenberg, 51–70. Westport, CT: Praeger.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2006. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." *Educ. Finance and Policy* 1 (2): 176–216.

Carrell, Scott E., Richard L. Fullerton, and James E. West. 2009. "Does Your Cohort Matter? Estimating Peer Effects in College Achievement." *J. Labor Econ.* 27 (3): 439–64.

Carrell, Scott E., and James E. West. 2008. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." Working Paper no. 14081, NBER, Cambridge, MA. http://www.nber.org/papers/w14081.

Cavalluzzo, Linda C. 2004. "Is National Board Certification an Effective Signal of Teacher Quality?" Technical Report no. 11204, CNA Corp., Alexandria, VA. http://www.cna.org/documents/CavaluzzoStudy.pdf.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *J. Human Resources* 41 (4): 778–820.

———. 2007. "How and Why Do Teacher Credentials Matter for Student Achievement?" Working Paper no. 12828, NBER, Cambridge, MA. http://ideas.repec.org/p/nbr/nberwo/12828.html.

Dee, Thomas S. 2004. "Teachers, Race, and Student Achievement in a Randomized Experiment." *Rev. Econ. and Statis.* 86 (1): 195–210. http://www.mitpressjournals.org/doi/abs/10.1162/003465304323023750.

———. 2005. "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *A.E.R. Papers and Proc.* 95 (2): 158–65.

Ehrenberg, Ronald G., and Dominic J. Brewer. 1994. "Do School and Teacher Characteristics Matter? Evidence from High School and Beyond." *Econ. Educ. Rev.* 13 (1): 1–17. http://ideas.repec.org/a/eee/ecoedu/v13y1994i1p1-17.html.

Ehrenberg, Ronald G., and Liang Zhang. 2005. "Do Tenured and Tenure-Track Faculty Matter?" *J. Human Resources* 40 (3): 647–59.

Ferguson, Ronald F., and Helen F. Ladd. 1996. "How and Why Money Matters: An Analysis of Alabama Schools." In *Holding Schools Accountable: Performance-Based Reform in Education*, edited by Helen F. Ladd, 265–98. Washington, DC: Brookings Inst. Press.

Figlio, David N., and Lawrence W. Kenny. 2007. "Individual Teacher Incentives and Student Performance." *J. Public Econ.* 91 (5–6): 901–14.

Goldhaber, Dan, and Emily Anthony. 2007. "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Rev. Econ. and Statis.* 89 (1):134–50. http://www.mitpressjournals.org/doi/abs/10.1162/rest.89.1.134.

Good, Phillip I. 2006. *Resampling Methods: A Practical Guide to Data Analysis.* 3rd ed. Boston: Birkhauser.

Hanushek, Eric A. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *A.E.R. Papers and Proc.* 61 (2): 280–88.

Harris, Douglas N., and Tim R. Sass. 2007. "Teacher Training, Teacher Quality and Student Achievement." Research Publications, Teacher Quality Research, Tallahassee, FL. http://www.teacherqualityresearch.org/teacher_training.pdf.

Hoffmann, Florian, and Philip Oreopoulos. 2009. "Professor Qualities and Student Achievement." *Rev. Econ. and Statis.* 91 (1): 83–92. http://www.mitpressjournals.org/doi/abs/10.1162/rest.91.1.83.

Jacob, Brian A., and Lars Lefgren. 2004. "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago." *J. Human Resources* 39 (1): 50–79.

Jacob, Brian A., Lars Lefgren, and David Sims. 2010. "The Persistence of Teacher-Induced Learning Gains." *J. Human Resources*, forthcoming.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." *Econ. Educ. Rev.* 27 (6): 615–31.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper no. 14607, NBER, Cambridge, MA. http://www.nber.org/papers/w14607.

Koerner, James D. 1963. *The Miseducation of American Teachers.* Boston: Houghton Mifflin.

Lehmann, E. L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* 3rd ed. Texts in Statistics. Secaucus, NJ: Springer.

McCaffrey, Daniel J., J. R. Lockwood, Daniel Koretz, and Laura Hamilton. 2004. *Evaluating Value-Added Models for Teacher Accountability.* Monograph no. 158. Santa Monica, CA: Rand Corp.

Murnane, Richard. 1975. *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, MA: Ballinger.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata*. 2nd ed. College Station, TX: Stata Press.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models.* 2nd ed. Thousand Oaks, CA: Sage.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools and Academic Achievement.” *Econometrica* 73 (2): 417–58.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *A.E.R. Papers and Proc.* 94 (2): 247–52.

Rothstein, Jesse. 2010. “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement.” *Q.J.E.* 125 (1): 175–214.

Summers, Anita A., and Barbara L. Wolfe. 1977. “Do Schools Make a Difference?” *A.E.R.* 67 (4): 639–52.

Vandevoort, Leslie G., Audrey Amrein-Beardsley, and David Berliner. 2004. “National Board Certified Teachers and Their Students’ Achievement.” *Educ. Policy Analysis Archives* 12 (46).

Weinberg, Bruce A., Masanori Hashimoto, and Belton M. Fleisher. 2009. “Evaluating Teaching in Higher Education.” *J. Econ. Educ.* 40 (3): 227–61.