TIME TO RAISE QUESTIONS ABOUT STUDENT RATINGS

*Linda B. Nilson
Clemson University*

Students have changed a great deal in the past few decades, but has the validity of student ratings correspondingly changed with them? Drawing on recent research conducted on students, this chapter examines the relationship between student ratings and student learning, the biases found in these ratings, and their factual accuracy and apparent truthfulness. It also addresses why findings of early and recent studies differ, what information student ratings now provide, and how institutions and faculty have consolidated the effects on ratings that today's students initiated.

2

The bulk of research on validity of student ratings and their contaminating biases was conducted in the 1970s and 1980s, when a very different generation of students filled classrooms. Dozens of publications on the millennial generation document how much students have changed since then (for example, Bauerlein, 2008; Lancaster & Stillman, 2003; Nathan, 2005; Pryor, Hurtado, DeAngelo, Palucki Blake, & Tran, 2011; Singleton-Jackson, Jackson, & Reinhardt, 2010; Twenge, 2007). They tell us that today's students under thirty years old devalue academics and reflection; feel entitled to higher education, high grades, top-quality customer service (as they define it); and attend college for primarily instrumental reasons—that is, not for the sake of learning but for the credential that will get them a job.

Furthermore, given that almost half of these students easily graduated from high school with an A-average, they come to college with an inflated estimation of their abilities and knowledge. They “earned” these A’s with little effort; in fact, nearly two-thirds of entering freshmen in 2009 devoted less than six hours a week to their homework as seniors in high school—a proportion of entering freshmen that has been increasing steadily since 1987 (Pryor et al., 2011). Being so unaccustomed to working hard and practicing persistence in their learning, these students do not study much more in college (Babcock & Marks, 2010; Pryor et al., 2011). Besides, they do not think they should have to because, in their fixed intelligence mind-set, they are “smart.” Aside from misrepresenting the way the brain actually works, this mind-set interferes with learning by inducing students to avoid challenges, give up easily, ignore criticism, and value high grades above mastery (Dweck, 2006). Yet, achievable challenge, time on task, and high expectations of performance are associated with greater student learning.

Given the values, beliefs, and behaviors of today’s students, it is time to raise questions about the validity of student ratings—specifically to reexamine the key assumptions and older literature. Fortunately, new research studies are available, and this chapter synthesizes their findings to identify what these ratings are really measuring.

What Validity Means for Student Ratings

The validity of student ratings as indicators of teaching effectiveness has rested on two findings and one assumption.

The first of the two findings is the moderately strong positive relationship between student ratings and student learning and achievement. Cohen (1981, p. 281) stated it very clearly: “It [teaching effectiveness] can be further operationalized as the amount students learn in a particular course. . . . If student ratings are to have any utility in evaluating teaching, they must show at least a moderately strong relationship to this index.” His meta-analysis found correlations between global rating items and learning, as measured by an external exam, of between .44 and .47. Meta-analyses by Marsh (1984) and Feldman (1989) synthesized comparable results and found a few slightly higher correlations between certain criterion-specific items and learning. These two literature reviews made convincing cases by focusing on studies of multisection courses with common syllabi and exams. Together, these three meta-analyses provided the justification for using student ratings as a proxy for student learning in faculty reviews.

The second finding is the minimal effect of rating biases due to extraneous variables. Cashin's (1995) meta-analysis identified a short list of biases: students' prior interest in the course subject matter, their expected grades, the level of course, the academic field, the purpose of ratings, the instructor's presence or absence while students filled out rating forms, and the anonymity of the ratings. But none of these except students' prior interest had notable effects, and the correlation between expected grades and ratings was largely mediated by student learning and motivation. In addition, a whole host of variables had no significant impact on ratings: students' gender, age, level, grade point average, and personality; instructor's gender, age, teaching experience, race, personality, and research productivity; class size; time of day of class meetings; and time in the semester that students filled out the rating forms. Cashin maintained that the effects of instructor rank and expressiveness were not biasing because they were related to learning. Although the findings varied some across studies, Hoyt (1997) estimated that biases accounted for only between 5 and 25 percent of the variance in rating items.

The assumption was the presumed honesty of students when filling out rating forms. So strong was this assumption that it attracted little research until the 1990s.

Recent Findings on Student Ratings

The research that substantiates the validity of student ratings is hardly recent. In fact, the main validity argument—the relationship of student ratings to learning—rests on the oldest data. All three of the key meta-analyses cited above were based on data collected in the 1970s, well over thirty years ago, and published on or before 1980. Feldman's later publications on the subject (1997, 1998, 2007) drew from the same aging data set. Cashin's (1995) literature review contained no studies published after 1994. Perhaps the time factor would not matter if the education-relevant values, attitudes, and behaviors of the raters had not changed, but they have, and quite radically. This is why only the most recent studies will answer the questions raised.

Are Student Ratings Still Related to Learning?

No. In fact, the relationship may be negative. Recent research has shifted the measure of learning from students' exam performance at the end of the course to their performance (grade) in follow-up courses. This latter measurement taps longer-lasting learning, which is presumably deeper

learning. While student ratings on multiple items (global and criterion-specific) and students' grades were positively correlated in the base course, this was not the case in follow-up courses, according to studies conducted at the U.S. Air Force Academy and Duke University (Carrell & West, 2010; Johnson, 2003). Ratings of the base course instructor and students' grades in follow-up courses were negatively correlated. In other words, the more effective introductory instructors—those who best prepared their students for advanced courses in the discipline—received lower ratings. In another study, this one conducted at Ohio State, student ratings on multiple items bore no statistically significant relationship to learning once grades, which did correlate with ratings, were controlled (Weinberg, Hashimoto, & Fleisher, 2009). In fact, in his recent meta-analysis, Clayson (2009) could not locate a single study documenting a positive relationship between student learning and student ratings that was published after 1990. A study at the University of California, Riverside, uncovered a weak relationship between student ratings and learning (Marks, Fairris, & Beleche, 2010), but the learning measured was not of the long-term type. Rather it was based on students' scores on a high-stakes final exam administered across multiple sections.

Might ratings be related to students' perceived learning? Indeed, two of the studies that found no relationship between objective measures of student learning and student ratings did find a positive one between students' perceived learning and their ratings (Johnson, 2003; Weinberg et al., 2009). But perceived learning is no substitute for actual learning. Weinberg et al. (2009) did not find any relationship between the two. Nor did the Wabash Study (Bowman, 2011), which encompasses seventeen institutions. Bowman examined how accurately students' perceived learning progress mirrored their actual progress on several longitudinal assessments—a series of objective tests, the Collegiate Assessment of Academic Proficiency focusing on critical and analytical reasoning, and the Defining Issues Test-2 assessing moral reasoning—and he reported no significant relationships. In fact, quite a few studies have found that students tend to overestimate their knowledge and abilities, except possibly the best students (Bell & Volckmann, 2011; Kruger & Dunning, 1999, 2002; Longhurst & Norton, 1997; Miller & Geraci, 2011).

Are Student Ratings More Biased Than in the Past?

Yes. Research published in the past decade confirms and adds to Cashin's short list of biases. The newest contaminants include well over a dozen variables extraneous to learning and largely beyond faculty

control: instructor's charisma (Shevlin, Banyard, Davies, & Griffiths, 2000), physical attractiveness (Hamermesh & Parker, 2005), personality (congeniality, confidence, optimism, and enthusiasm, estimated by studies cited in Clayson, 2011, to explain 50 to 75 percent of the variance in student ratings), age (McPherson & Jewell, 2007; Zabaleta, 2007), rank (Isley & Singh, 2007; McPherson & Jewell, 2007), gender in the sciences and engineering (Potvin, Harazi, Tai, & Sadler, 2009; Superson, 1999); membership in a disadvantaged racial group (McPherson & Jewell, 2007) and Asian accent (Gravois, 2005); the length of class meeting, in that longer meetings lower ratings (Isley & Singh, 2007; McPherson & Jewell, 2007); the timing of the ratings, in that collecting the data shortly before tests and due dates of major assignments and after returning graded work depresses ratings (Hall & Fitzgerald, 1995); class size (McPherson & Jewell, 2007; Wines & Lau, 2006); the number of rows in the classroom (Wines & Lau, 2006); other aspects of the classroom, the quality of the curriculum, and the functionality of the room's technology (Nowell, Gale, & Handley, 2010); and the students' perceived or anticipated grade in the course (Clayson, Frost, & Sheffet, 2006; Isley & Singh, 2007; McPherson & Jewell, 2007; Wines & Lau, 2006; Zabaleta, 2007).

Regarding the last bias, the recent studies do not attribute the relationship between expected grade and ratings to greater student learning or motivation, as have Cashin (1995) and others. To the contrary, student scores on items tapping course rigor, challenge of the content, and required student effort—all presumably related to learning—vary negatively with instructor ratings (Centra, 2003; Steiner, Holley, Gerdes, & Campbell, 2006; Weinberg, Fleisher, & Hashimoto, 2007; others cited in Clayson, 2009). Clayson (2011) hypothesized six possible explanations for the grade-ratings correlation: teaching effectiveness/learning, student motivation, grade attribution, prior student and instructor characteristics, grading leniency, and reciprocity. His prior research (Clayson et al., 2006) led him to endorse the reciprocity theory over the others. It posits that a student's evaluation of an instructor simply reflects what the student sees to be the instructor's evaluation of her. In other words, students give what they see themselves receiving. In addition, the correlation between expected grade and ratings has increased over the past few decades; it used to be .10 to .30 but is now .45 to .50 (Clayson, 2011). So what Cashin (1995) dismissed as a mild bias has become a major predictor of student ratings.

All of these findings—those documenting more and stronger biases and those suggesting that rigor, challenge, and required effort lower student

ratings—add to the evidence that learning and ratings are either uncorrelated or negatively correlated.

Are Student Ratings Honest?

We have assumed that students generally are telling the truth, as they see it, when they fill out the rating forms. Yet many of us have also found factual inaccuracies in our own student ratings and written comments and, as faculty developers, have counseled other instructors who have. We have said and heard protests like these: “But I’ve *never* been late to class!” “But I returned all the papers and exams within a week!” “But I followed the syllabus faithfully!”

Surprisingly few studies have addressed this problem. Stanfel (1995) quizzed his students on their knowledge of his evaluation procedures in his course, and all of them correctly described the procedures. In addition, whenever he returned graded work, he asked them to sign a document stating that they had received their graded work at the first possible opportunity. Yet on the end-of-semester rating forms, only 3 percent of his students “strongly agreed” with the item, “The instructor explains clearly to students how they are evaluated,” and 64 percent “disagreed” or “strongly disagreed.” In response to the item, “Tests and written assignments are graded and returned in a reasonable period of time,” only 3 percent “strongly agreed” while over 46 percent “disagreed” or “strongly disagreed.” In a similar study, Sproule (2002) diligently returned all graded work for one of his classes the very next class meeting during the entire semester, but only half of his students’ responses to the relevant rating item reflected this.

Were these misrepresentations of the truth due to students’ forgetting, misunderstanding, or lying? Clayson and Haley (2011) surveyed students about their honesty in their ratings and written comments, and the disturbing results confirmed Stanfel’s and Sproule’s worst suspicions: about one-third of the students confessed to “stretching the truth,” 56 percent said they knew peers who had, and 20 percent admitted to lying in their comments. Moreover, half the students did not think that what they did constituted a kind of cheating.

While few in number, Stanfel’s, Sproule’s, and Clayson and Haley’s studies cast doubt on the validity of all student responses on rating forms. When added to the recent research documenting serious biases in the ratings and a disconnect between student learning and the ratings, they lead us to conclude that student ratings lack content validity and all three kinds of construct validity—convergent, discriminant, and divergent (Clayson, 2011).

Older Studies in Perspective

How did the studies of student ratings conducted in the 1970s and 1980s obtain the results that they did? Of course, they examined only short-term learning using end-of-course final exams. But Marks et al. (2010) did as well and found only a weak association between learning and ratings. Most likely the explanation lies with the respondents, because the early ones had very different values, attitudes, and behaviors from today's students. This earlier student generation attended college more for learning-focused and life-enhancement reasons than for good jobs. In 1967, 86 percent of freshmen said they were attending to develop a meaningful philosophy of life (Hong, 2004). Today about as high a proportion say "to get a better job," and less than 40 percent care about developing a meaningful philosophy of life (Pryor et al., 2011).

In the past, what students wanted out of college more closely matched the mission of higher education—liberal arts education, in particular. In addition, the K–12 system in the United States apparently operated more effectively, as its students scored high on reading and math in international tests. Entering college students either had or had to quickly acquire solid, independent study skills and habits, or institutions flunked them out. (High freshman attrition rates were points of pride for some universities.) It seems that these students realized that their learning was their responsibility. What they did not glean from the lectures, they had to learn on their own or seek help from faculty or teaching assistants. (Few, if any, academic assistance centers existed then.) Given their reasons for attending college and their educational experience, these students probably expected rigor, valued challenge, and gave higher ratings to professors from whom they learned the most. Maybe a combination of student learning and motivation did mediate the relationship between grades and ratings at the time.

In addition, institutions did not see themselves operating within the customer-seller model. They functioned more within the client-professional model, exemplified by the patient-physician or the client-attorney relationship. It is an unequal pairing in that both parties agree that the professional has expertise that the client needs to draw on. While clients want professionals to display caring and warmth, what they most want—and are paying for—is well-informed advice and appropriate follow-up action. For example, a patient expects to receive appropriate testing, an accurate diagnosis, proper treatment, and, if possible, restoration to health. In the lawyer's office, a client wants an expert assessment of her situation, followed by legal actions that promote and protect her best interests. Being

a reasonable patient or client entails accepting bad news, taking prescribed actions, and enduring side effects, sacrifices, or costs if they serve her longer-term, rational interests.

When faculty assumed the role of professional, student clients came to college respecting their expertise and not anticipating a democratic, empowering classroom experience. While they no doubt valued an instructor's caring, warmth, and ability to motivate them, they expected her to have up-to-date knowledge and skills and hoped that she could convey them effectively. They might have had to work very hard, accept critical feedback, and deal with some low grades to get through their learning process, but these were par for the course. While this description of the professor-student relationship reflects the pure client-professional model, it may accurately represent the norms of the time.

Institutional and Faculty Responses to Today's Students

Nuhfer (2010) interprets student ratings on global items as reflections of student affect toward an instructor and valid measures of student satisfaction with the instructor and course experience. This makes sense given how strongly instructor personality and expected grade influence these ratings. Today's students want an instructor they can relate to, who is expressive and energetic, and who cares about and empathizes with them (Chonko, 2004; Clayson, 2011; Kelley, Conant, & Smart, 1991; La Lopa, 2011; Walsh & Maffei, 1994). If an instructor's ability to project such a persona motivated students to learn more, then ratings and learning would be positively related, but we have already established that they are not. Students also want the good grades that they were accustomed to getting before college (Pryor et al., 2011) and to preserve their positive self-concept. Thus, they accordingly reward faculty who give them high grades with high ratings.

In general, colleges and universities have come to view themselves as sellers in the market economy, competing to attract students and striving to retain them. In this marketplace model, student satisfaction means customer satisfaction, whether the customer is the student or her parents or both. (If a dependent student is unhappy about something, the parents typically know about it and share that unhappiness, and they often contact the instructor or administrators to get satisfaction.) According to Kirp (2005), colleges and universities have dutifully responded to what their prospective and current customers want, upgrading residence halls, building recreational facilities, improving student services, and making academics more "palatable" to the general public. Although no definitive

study shows that institutions have dumbed down their curricula, Babcock and Marks (2011) uncover evidence that the standards of effort that they set for students have dropped precipitously since 1961, and students have allocated their out-of-class time toward social and leisure activities. At the same time, colleges and universities have allowed considerable grade inflation over the past thirty years (Johnson 2003; Rojstaczer, 2009). They have also placed increasing importance on student ratings in hiring, retaining, and promoting faculty—in particular, the global ratings that Nuhfer interprets as an affective satisfaction measure.

The institutional response to market pressures has changed faculty incentives and, in turn, faculty behavior (Sperber, 2005). To keep up their student ratings, instructors have generally graded more leniently than in the past (thus, grade inflation) and have reduced course workloads (Johnson, 2003; Sperber, 2005). Some even may have learned to practice the social immediacies that make students feel cared for and respected. Williams and Ceci (1997) demonstrated that faculty could raise their ratings just by learning to act more enthusiastic, but how many instructors have adopted this strategy is unknown. (Interestingly, Ceci's enhanced ratings did not result in better student performance on the final exam.) Conversely, upholding high standards seems to yield no institutional rewards and can result in lower student ratings and the accompanying career consequences (Babcock & Marks, 2011).

In view of institutional responses to the values, attitudes, and behaviors of today's students, followed by faculty responses to the resulting new incentives, the diminishing validity of student ratings is due to changes not only in students but also in colleges and universities and their faculties. Students may have initiated the trend, but they have received a great deal of cooperation. To reverse the trend, institutions would have to be willing to sacrifice student satisfaction for greater learning and accept the possibility of higher attrition. They would also have to deemphasize (if not eliminate) student ratings in faculty reviews and rely largely on well-informed peer review and measures of student learning. Instructors would then be free to challenge students more and work them harder without fearing that falling student ratings would derail their academic career. Faculty would also have the discretion to fail indifferent and poorly performing students without parental interference or administrative sanction. As a result, students might learn more in college than they currently do, and some of those who failed might return to college later, ready to work and learn.

The current accreditation system is not designed to increase rigor and counterbalance the power of student ratings. Accrediting agencies require

institutions to develop and assess student learning outcomes, but they do not require that all graduating students achieve those outcomes at the target level. All they do demand is that institutions try new strategies that promise to improve the results. So far, these strategies have left a lot to be desired.

Conclusion

Many colleges and universities go through cycles with their student rating forms, every several years reexamining the items they settled on at the end of the previous cycle. They often look to faculty developers to provide the scholarship and guidance to improve these forms and decide how they should be used. Developers typically recommend adding evidence of teaching effectiveness beyond student ratings, such as peer review of course materials, well-informed classroom observations, and pre- and posttests of learning (Arreola, 2007; Stark-Wroblewski, Ahlering, & Brill, 2007). Still, the primacy of student ratings persists. It is ironic that these ratings have acquired increasing importance in tenure, promotion, and reappointment decisions over the same time period that their validity has waned. The question never raised is whether we should use these ratings in faculty reviews at all (faculty private use is not problematic). It is time to initiate the discussion.

The academy has to decide what “teaching effectiveness” really means. If it means student learning, as Cashin (1988) defined it, then institutions should assess it using measures that reflect learning, not student ratings. These ratings belong in the faculty review process only to the extent that faculty and administrators regard student/customer satisfaction as an important goal unto itself, even though it has nothing to do with and can actually work against student learning. Deciding the fate of instructors based in whole or in part on their student ratings is an institution’s choice, but it should inform its faculty explicitly that they are being assessed on student satisfaction, independent of learning, if that is the case.

REFERENCES

- Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system*. San Francisco, CA: Jossey-Bass/Anker.
- Babcock, P., & Marks, M. (2010, August). *Leisure College, USA: The decline in student study time*. (American Enterprise Institute for Public Policy Research, No. 7). Retrieved from <http://www.aei.org/outlook/100980>

- Babcock, P. S., & Marks, M. (2011). The falling time cost of college: Evidence from a half century of time use data. *Review of Economics and Statistics*, 93, 468–478.
- Bauerlein, M. (2008). *The dumbest generation: How the digital age stupefies young American and jeopardizes our future*. New York, NY: Tarcher/Penguin.
- Bell, P., & Volckmann, D. (2011). Knowledge surveys in general chemistry: Confidence, overconfidence, and performance. *Journal of Chemical Education*, 88(11), 1469–1476.
- Bowman, N. A. (2011, April 11). *The validity of college seniors' self-reported gains as a proxy for longitudinal growth*. Paper presented at the annual meetings of the American Educational Research Association, New Orleans, LA.
- Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3), 409–432.
- Cashin, W. E. (1995) *Student ratings of teaching: The research revisited*. (IDEA Paper No. 32). Manhattan: Center for Faculty Development and Evaluation, Kansas State University.
- Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research* (IDEA Paper No. 20.). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.
- Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less coursework? *Research in Higher Education*, 44, 495–518.
- Chonko, L. B. (2004). If it walks like a duck....: Concerns about quackery in marketing education. *Journal of Marketing Education*, 26, 4–16.
- Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, 31(1), 16–30. Retrieved from <http://jmd.sagepub.com/content/31/1/16.full.pdf+html>
- Clayson, D. E. (2011). *A multi-disciplined review of the student teacher evaluation process*. Retrieved from <http://business.uni.edu/clayson/Ext/SETSummary2011.doc>
- Clayson, D. E., Frost, T. E., & Sheffet, M. J. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning and Education*, 5(1), 52–65.
- Clayson, D. E., & Haley, D. A. (2011). Are students telling us the truth? A critical look at the student evaluation of teaching. *Marketing Education Review*, 21, 101–112.
- Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisession validity studies. *Review of Educational Research*, 51, 281–309.

- Dweck, C. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.
- Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, 30, 583–645.
- Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368–395). New York, NY: Agathon Press.
- Feldman, K. A. (1998). Identifying exemplary teachers and teaching: evidence from student ratings. In K. A. Feldman & M. B. Paulsen (Eds.), *Teaching and learning in the college classroom* (2nd ed., pp. 391–414). New York, NY: Simon & Schuster.
- Feldman, K. A. (2007). Identifying exemplary teachers and teaching: evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based approach* (pp. 93–129). New York, NY: Springer.
- Gravois, J. (2005, April 8). Teach impediment: When the students can't understand the instructor, who is to blame? *Chronicle of Higher Education*. Retrieved from <http://chronicle.com/article/Teach-Impediment/33613/>
- Hall, C., & Fitzgerald, C. (1995). Student summative evaluation of teaching: Coder of practice. *Assessment and Evaluation in Higher Education*, 20(3), 307–311.
- Hamermesh, D. S., & Parker, A. M. (2005). Beauty in the classroom: Professorial pulchritude and putative pedagogical productivity. *Economics of Education Review*, 24, 369–376.
- Hong, P. Y. (2004, January 26). Money top goal of college freshmen. *Los Angeles Times*. Retrieved from <http://articles.latimes.com/2004/jan/26/local/me-survey26>
- Hoover, E. (2009, October 11). The millennial muddle. *Chronicle of Higher Education*. Retrieved from <http://chronicle.com/article/The-Millennial-Muddle-How/48772/>
- Hoyt, D. P. (1997). *Studies of the impact of extraneous variables*. Manhattan: IDEA Center, Center for Faculty Development and Evaluation, Kansas State University. Retrieved from <http://www.bus.lsu.edu/accounting/faculty/lcrumbley/idea.html>
- Isley, P., & Singh, H. (2007). Does faculty rank influence student teaching evaluations? Implications for assessing instructor effectiveness. *Business Education Digest*, 16, 47–59.

- Johnson, V. E. (2003). *Grade inflation: A crisis in higher education*. New York, NY: Springer.
- Kelley, C. A., Conant, J. S., & Smart, D. T. (1991). Master teaching revisited: Pursuing excellence from the students' perspective. *Journal of Marketing Education*, 13, 1–10.
- Kirp, D. L. (2005). This little student went to market. In R. H. Hersch & J. Merrow (Eds.), *Declining by degree: Higher education at risk* (pp. 113–130). New York, NY: Palgrave Macmillan.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121–1134.
- Kruger, J., & Dunning, D. (2002). Unskilled and unaware—but why? *Journal of Personality and Social Psychology*, 82(2), 189–192.
- La Lopa, J. M. (2011). Student reflection on quality teaching and how to assess it in higher education. *Journal of Culinary Science and Technology*, 9(4), 282–292.
- Lancaster, L. C., & Stillman, D. (2003). *When generations collide: Who they are. Why they clash. How to solve the generation puzzle at work*. New York, NY: HarperCollins.
- Longhurst, N., & Norton, L. S. (1997). Self-assessment in coursework essays. *Studies in Educational Evaluation*, 23(4), 319–330.
- Marks, M., Fairris, D., & Beleche, T. (2010, June 3). *Do course evaluations reflect student learning? Evidence from a pre-test/post-test setting*. Riverside: Department of Economics, University of California, Riverside. Retrieved from <http://faculty.ucr.edu/~mmarks/Papers/marks2010course.pdf>
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76, 707–754.
- McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, 88(3), 868–881.
- Miller, T. M., & Geraci, L. (2011, January 24). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/a0021802.
- Nathan, R. (2005). *My freshman year. What a professor learned by becoming a student*. Ithaca, NY: Cornell University Press.
- Nowell, C., Gale, L. R., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment and Evaluation in Higher Education*, 35(4), 463–475.

- Nuhfer, E. B. (2010). *A fractal thinker looks at student ratings*. Retrieved from <http://profcamp.tripod.com/fractalevals10.pdf>
- Potvin, G., Hazari, Z., Tai, R. H., & Sadler, P. (2009). Unraveling bias from student evaluations of their high-school teachers. *Science Education*, 93(5), 827–845. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/sce.20332/pdf>
- Pryor, J. H., Hurtado, S., DeAngelo, L., Palucki Blake, L., & Tran, S. (2011). *The American freshman: National norms fall 2010*. Los Angeles, CA: Higher Education Research Institute, UCLA.
- Rojstaczer, S. (2009). *GradeInflation.com: Grade inflation at American colleges and universities*. Retrieved from <http://gradeinflation.com/>
- Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment and Evaluation in Higher Education*, 25, 397–405.
- Singleton-Jackson, J. A., Jackson, D. L., & Reinhardt, J. (2010). Students as consumers of knowledge: Are they buying what we're selling? *Innovative Higher Education*, 35(5), 343–358.
- Sperber, M. (2005). How undergraduate education became college lite—and a personal apology. In R. H. Hersch & J. Merrow (Eds.), *Declining by degree: Higher education at risk* (pp. 131–144). New York, NY: Palgrave Macmillan.
- Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review*, 21, 287–295.
- Stanfel, L. E. (1995). Measuring the accuracy of student evaluations of teaching. *Journal of Instructional Psychology*, 22(2), 117–125.
- 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 and Evaluation in Higher Education*, 44(5), 539–556. Retrieved from <http://www.tandfonline.com/doi/pdf/10.1080/02602930600898536>
- Steiner, S., Holley, L. C., Gerdes, K., & Campbell, H. E. (2006). Evaluating teaching: Listening to students while acknowledging bias. *Journal of Social Work Education*, 42, 355–376.
- Superson, A. M. (1999). Sexism in the classroom: The role of gender stereotypes in the evaluation of female faculty. *APA Newsletter on Feminism and Philosophy*, 99(1), 46–51.
- Twenge, J. M. (2007). *Generation me: Why today's young Americans are more confident, assertive, entitled and more miserable than ever before*. New York, NY: Free Press.

- Walsh, D. J., & Maffei, M. J. (1994). Never in a class by themselves: An examination of behaviors affecting the student-professor relationship. *Journal of Excellence in College Teaching*, 5(2), 23-49.
- Weinberg, B. A., Fleisher, B. M., & Hashimoto, M. (2007). *Evaluating methods of evaluating instruction: The case of higher education*. (NBER Working Paper No. 12844). Retrieved from <http://www.nber.org/papers/w12844>.
- Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *Journal of Economic Education*, 40(3), 227-261.
- Williams, W. M., & Ceci, S. J. (1997, September/October). How'm I doing? Problems with student ratings of instructors and courses. *Change*, 29(5), 13-23.
- Wines, W. A., & Lau, T. J. (2006). Observations on the folly of using student evaluations of college teaching for faculty evaluation, pay, and retention decisions and its implications for academic freedom. *William and Mary Journal of Women and the Law*, 13(1). Retrieved from <http://scholarship.law.wm.edu/cgi/viewcontent.cgi?article=1089&context=wmjowl>
- Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, 12(1), 55-76.