ACUE

STUDENT EVALUATIONS OF TEACHING:

BEST PRACTICES

Meghan Snow, EdD
Elizabeth K. Lawner, PhD
Jonathan Gyurko, PhD
Theo Pippins, PhD

August 2022

ACUE

THIS BRIEF PROVIDES

- ✓ Context on why student evaluations of teaching, rather than generically of courses, are increasingly important
- ✓ Best practices to reduce bias, ensure relevance, and make survey purposes clear to students and faculty
- ✓ Discussion of the bias-free student survey developed by the Association of College and University Educators (ACUE) and in use at hundreds of colleges and universities nationwide

Introduction

Quality instruction is increasingly recognized as essential to student success and equity in higher education. For example, a survey of 30,000 college graduates for the Gallup-Purdue Index found that students were more likely to be in good jobs and leading fulfilled lives as a result of effective teaching (Gallup & Strada Education Network, 2018). The National Survey of Student Engagement (NSSE) finds a positive correlation between more effective teaching and the degree to which students feel valued by their institution (2022). Research by the Association of College and University Educators (ACUE) shows stronger achievement and closed or narrowed equity gaps among students taught by faculty who are certified in the use of evidence-based practices (Lawner & Snow, 2020).

Student evaluations of teaching—when conducted properly—can indicate if teaching is evidence-based and likely to improve academic achievement and equity.

Based on this increasing recognition, philanthropic support for faculty development is growing and tops the list of charitable priorities, notes *The Chronicle of Higher Education* (Blumenstyk, 2022). Gates, Carnegie, Strada, and other foundations have given multimillion-dollar grants to strengthen collegiate instruction. Higher education journalist Paul Fain recently summarized, "Perhaps the biggest factor in student success—including after college—is personal engagement with a faculty member" (2022, Credentials for Instructors section).

These investments indicate that effective teaching is a "student success" intervention. But it's fair to ask, "How do we know?" That is, to what extent is such investment by institutions and foundations leading to actual changes in teaching, with college educators incorporating approaches shown to promote student engagement, persistence, and deeper learning?

ACUE studies show that the best way to examine the effects of professional development on changes in teaching and consequent positive impact with students is through rigorous and longitudinal studies. This approach compares outcomes among students taught by faculty who are certified in effective instructional practices to those not yet certified (Gyurko & Snow, 2020). Such robust studies, on top of decades of research on the positive effects of good teaching, should give faculty, administrators, and policymakers confidence in the value of teaching to learning, degree completion, and career-readiness.

But college and universities leaders cannot wait the years it takes to conduct such evaluations, given the real-time demands they face to improve graduation rates and close equity gaps. Instead, they must rely on extant findings generalizable to their context. Closer to home, they should track valid leading indicators, such as student evaluations of teaching. If conducted properly, by removing normative judgements of quality and bias in instrumentality, such surveys provide reliable information regarding the quality of teaching, on which leaders can make decisions regarding policies, practices, and resource allocation (MacCormack et al., 2018).

This was funded by the Bill & Melinda Gates Foundation. The views expressed are those of the author(s) and should not be attributed to the funder.

Student Evaluation of Teaching

Student evaluations are ubiquitous in higher education (Berk, 2005). But we draw a distinction between evaluations of *courses,* which purport to examine a range of issues, and evaluations of *teaching,* which are specific to instructional practices and learning conditions. Such teaching evaluations are sometimes administered for formative feedback (Berk, 2005), providing professors with information about their teaching strengths and areas in need of improvement.

More frequently, student evaluations of teaching (SETs) and course evaluations are used as summative measures of an instructor's effectiveness and attached to high-stakes employment decisions including promotion, tenure, and, in the case of contingent professors, contract renewal. But widespread dissatisfaction about the quality and validity of these instruments rightly challenges their use in such situations (e.g., Spooren et al., 2013).

With SETs in such common use, particularly in ways that have material consequences for faculty, and the growing importance of teaching to student success and equity, there are best practices that higher education should follow to ensure that these instruments are valid and useful.

Recommendations

1: Ensure that instruments are not biased against particular groups of faculty.

Research shows that student instructional ratings tend to be biased against faculty of color (Hamermesh & Parker, 2005; Huston, 2006), non-native English speakers, (Hamermesh & Parker, 2005), and women (Boring et al., 2016; Macnell et al., 2015), particularly in male-dominated disciplines (Huston, 2006). In practice, this means that certain groups of faculty are, on average, consistently rated differently than their peers and at-risk when survey results are used in high-stakes decisions. By comparison, a bias-free instrument would see any variance in ratings mimic the distribution of evaluations across all faculty members. Thus, it is important to routinely test and correct SETs for bias.

Routinely test and correct SETs for bias.

In testing for bias, be sure to account for factors that affect student ratings but are unrelated to teaching effectiveness. For example, evaluations tend to be lower for required courses (Huston, 2006), and course characteristics should be included in statistical models when examining potential biases. In addition, factors that legitimately impact teaching effectiveness should be accounted for to isolate differences that are due to bias rather than actual differences in instruction. For example, analyses that do not control for years of experience may seem to find evidence of bias against faculty of color, who are, on average, younger than the professoriate at large.

If analyses indicate that a SET instrument may be biased, colleges and universities should look for patterns in the survey items. The literature on bias in SETs (e.g., Peterson et al., 2019) and unconscious bias in other contexts can inform revisions. It is then necessary to test any new or revised items and instruments. One way to mitigate bias on SET is to remove general items and make questions as specific as possible, since bias tends to show up in more general items (Huston, 2006). Moreover, Reinsch and colleagues (2020) have suggested that the most problematic items should be removed from SET scales.

If revising a SET instrument is not feasible, an alternative is to adjust scores based on the estimated degree of bias, particularly if SETs are used in high-stakes employment decisions. Another option is to provide bias training to students before they complete SETs (Reinsch et al., 2020) or add information about bias to the instructions. Doing so has been found to reduce gender bias (Peterson et al., 2019). Even after a SET instrument or its administration has been improved to mitigate bias, colleges and universities should continue to routinely test the instruments to sustain collection of valid and reliable information.

2: Items should be written with the student in mind.

For SETs to be valid measures of teaching effectiveness, students need to be able to accurately respond to the questions. That means instructions and questions must use language that students understand and avoid jargon with which they may not be familiar. For example, students may not be able to rate whether "learning outcomes" were clear if they do not know what "learning outcomes" means. Instead, students can readily understand and indicate whether course "aims" or "assignment tasks" were clear. Wherever possible, plain language is best.

Best practice is to simply ask students what happened in class—what teaching practices were present or not, rather than their perceptions of how well the course ran.

Ensuring that students can accurately respond to SET items also means limiting questions to issues on which they have the perspective and ability to answer. For example, Carl Wieman asserts that "it is impossible for a student (or anyone else) to judge the effectiveness of an instructional practice except by comparing it with others that they have already experienced" (2015, p. 9). But such relative judgements tell nothing about quality against objective, professionally normed standards of practice.

Therefore, SETs should not ask students to *evaluate the effectiveness* of instruction, whether in general or for a specific technique. However, students are able to simply *describe whether a particular practice* was used and how often.

Wieman also cautions against having students report on how much they learned in a course, given the challenges of assessing one's own learning. Such questions are "sensitive to the level of expertise of the respondent" (Wieman, 2015, p. 9) and, again, relative to the respondents' experience rather than objective course expectations. Nor are students well positioned to evaluate their instructors' relative expertise in a content area, since students themselves are not experts.

This best practice remains: simply ask students *what* happened in class—what teaching practices were present or not, rather than their perceptions of *how well* the course ran. In other words, surveys should ascertain the extent to which effective teaching practices are used in the classroom by grounding the item development and interpretation of the data in the body of research about instruction in higher education, not by asking students outright if teaching was "effective".

3: SETs should have a clear purpose and be connected to outcomes of interest.

When colleges and universities use SETs, they should determine the purpose and demonstrate the connection to relevant outcomes. For example, if institutions want to assess student satisfaction, they should be clear on the reason, such as a possible connection between satisfaction and retention—and then they should assess whether their SET actually predicts retention. If the purpose of the SET is to provide formative feedback to faculty, results should not be used in punitive or evaluative ways, and items should be as specific as possible in order to provide actionable feedback. In those cases, results of SETs would also be more helpful for faculty if they were accompanied by resources and faculty development opportunities that faculty can use to improve specific areas assessed in the survey.

Determine whether SETs will be primarily formative or evaluative and provide professional learning opportunities to help faculty meet expectations.

If institutions plan to use SETs to evaluate faculty performance to inform employment decisions, they need to ensure that the instrument is measuring effective teaching and not just how much students like the faculty member. They can do so by analyzing whether their SET instrument is correlated with other validated measures of effective teaching or with measures of student engagement or success. Institutions might assume that SETs measure teaching effectiveness directly, or at least that students learn more from faculty who have higher ratings on such measures (Uttl et al., 2017). It is also true that some meta-analyses demonstrate relationships between student evaluations of teaching and student achievement (Cohen, 1980, 1981; Feldman, 1989). However, reanalysis of these studies showed that the prior findings were unduly influenced by studies with small sample sizes and publication bias, and actual relationships were not statistically significant (Uttl et al., 2017). This is not to say that no student evaluation instrument predicts achievement, rather that institutions should not automatically make that assumption about an instrument without testing its validity.

Finally, it is best practice to ensure that faculty have necessary training and support to teach well *before* SETs are used for any summative and high-stakes purpose, such as employment decisions. Until expectations of faculty are clear and professional learning opportunities are provided by the institution to help faculty meet these expectations, SETs should *only* be formative tools collecting data that can be used for improvement. Otherwise, faculty members are being held accountable for something they've not been given sufficient opportunity to develop.

ACUE's Survey for Student Evaluation of Teaching

ACUE developed a student survey in 2017 to provide faculty with real-time, formative feedback on their teaching. The survey is fully aligned to the 25 core teaching competencies and five major areas of practice defined in ACUE's nationally recognized Effective Practice Framework. As a result, faculty learning and implementing evidence-based practices through an ACUE course in effective teaching can receive information directly related to their development of best practices.

The survey's instructional practices scale comprises the bulk of the survey and is most similar to student evaluations of teaching. Students are asked to respond on a 5-point Likert scale to statements about their instructors' *use* of effective teaching practices. Importantly, each item is about a specific aspect of effective teaching, but never asks students to judge what is "effective." For example, instead of asking students whether instructors gave effective feedback, one item reads, "My instructor's feedback helped me to improve my performance in this course." Plus, items are written in plain language that students can understand and generate feedback that faculty can readily use to adjust their teaching.

The survey also includes measures of student self-efficacy and growth mindset, and background questions about the student demographics and the course. ACUE has refined the survey annually, including adding a measure of student belonging and expanding the instructional practices scale from 17 to 20 items.

At present, the survey is used by institutions engaged in strategic partnerships with ACUE. Faculty distribute their individualized survey links to their students, often through emails or announcements in their learning management system. Responses are completely anonymous and are reported on at the programmatic level for leadership decision-making, with faculty anonymized as well. In addition, faculty who receive sufficient responses to maintain anonymity of their students receive a personalized report for formative use.

To ensure validity and reliability, ACUE examines student responses for potential bias each semester. ACUE determines whether any systematic differences exist by faculty race/ethnicity and gender, after controlling for important faculty, student, and course characteristics (ACUE, 2022).

ACUE assumes the following: by controlling for factors that might affect students' ratings, such as their class year and whether the course meets requirements for their major or minor, as well as for factors that might impact faculty teaching effectiveness, such as their years of experience and number of effective teaching practices they have implemented, any remaining differences are likely due to bias, particularly if the same patterns occur across multiple administrations of the survey.

Bias Free

In the 2021-2022 academic year, analyses controlling for student, course, and faculty variables yielded no significant association between student ratings on the instructional practice scale and faculty race/ethnicity or gender, providing suggestive evidence that the instrument is free from racial and gender bias (ACUE, 2022).

This outcome was the result of several years of analysis and item revision, based on indications of suspected bias and informed by the best practices and extant research as detailed above. For instance, an item on grading was revised to remove the word "fair," since fair grading would mean a process carried out equitably for all students. Students do not witness the grading process and thus are not good judges of its fairness.

ACUE's revised survey items were piloted for at least one semester alongside the original items, with the order of the items randomized for each respondent. This tested whether the revisions did in fact reduce bias. Items found to be less biased then replaced their precursors on the instrument to form the improved survey that was tested during the 2021-2022 academic year. In total, four of the 20 items were revised and replaced. The process is ongoing, as ACUE will continue to test for bias and pilot new items.

Conclusion

Colleges and universities must ensure that student surveys are fair and accurate, whether for institutional research, formative feedback to faculty, or high-stakes employment decisions. Such accuracy—free of bias against particular groups of professors—serves a valuable purpose for decision-making, given that administrators are often dependent on leading indicators of change. Moreover, leaders should consider adopting ACUE's SET, as part of a program of faculty development, given its alignment with the recommendations presented here.

References

Association of College and University Educators. (2022, June). *ACUE student survey shows no evidence of bias.* [Research Brief 22]. https://acue.org/studentsurveybiasfree/

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48–62. https://www.isetl.org/ijtlhe/pdf/IJTLHE8.pdf

Blumenstyk, G. (2022, June 22). The edge: Your ideas for education philanthropists' next bets. *The Chronicle of Higher Education.* https://www.chronicle.com/newsletter/the-edge/2022-06-22

Boring, A., Ottoboni, K., & Stark, P. B. (2016, January 7). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research, 0*(0):1–11. DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Cohen, P. A. (1980). *A meta-analysis of the relationship between student ratings of instruction and student achievement.* [Unpublished doctoral dissertation]. The University of Michigan.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*(3), 281–309. http://dx.doi.org/10.2307/1170209.

Fain, P. (2022, June 23). The rise of outcomes-based loans. *Open Campus.* https://www.opencampusmedia.org/2022/06/23/the-rise-of-outcomes-based-loans/

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*(6), 583–645. https://doi.org/10.1007/BF00992392

Gallup & Strada Education Network. (2018). *2018 Strada–Gallup alumni survey: Mentoring college students to success.* Gallup. https://news.gallup.com/reports/244058/2018-strada-gallup-alumni-survey.aspx

Gyurko, J., & Snow, M. (2020). Our "directive": Quality teaching and learning. *Change: The Magazine of Higher Learning, 52*(5), 6–16. https://doi.org/10.1080/00091383.2020.1807873

Hamermesh, D. S., & Parker, A. (2005, August). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*(4), 369–376. http://dx.doi.org/10.1016/j.econedurev.2004.07.013

Huston, T. A. (2006, May). Race and gender bias in higher education: Could faculty course evaluations impede further progress toward parity? *Seattle Journal for Social Justice, 4*(2), 591–611. https://digitalcommons.law.seattleu.edu/sjsj/vol4/iss2/34

Lawner, E. K., & Snow, M. (2020, May 15). *Advancing academic equity at Broward College: Improved course completion and passing, particularly among Pell-eligible and Black students.* Association of College and University Educators. https://acue.org/wp-content/uploads/2020/06/ACUE_Broward-Tech-Report_100120.pdf

MacCormack, P., Snow, M., Gyurko, J., & Candio Sekel, J. (2018). *Connecting the dots: A proposed accountability method for evaluating the efficacy of faculty development and its impact on student outcomes.* Association of College and University Educators. https://acue.org/wp-content/uploads/2018/07/WP3_Connecting-the-Dots.pdf

Macnell, L., Driscoll, A., & Hunt, A. N. (2015, August). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303. https://doi.org/10.1007/s10755-014-9313-4

National Survey of Student Engagement. (2022). *Engagement insights: Survey findings on the quality of undergraduate education.* https://nsse.indiana.edu/research/annual-results/2021/index.html

Peterson, D. A. M., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe. K. (2019, May 15). Mitigating gender bias in student evaluations of teaching. *PLOS ONE, 14*(5), e0216241. https://doi.org/10.1371/journal.pone.0216241

Reinsch, R. W., Goltz, S. M., & Hietapelto, A. B. (2020). Student evaluations and the problem of implicit bias. *Journal of College and University Law, 1,* 114–139. https://digitalcommons.mtu.edu/michigantech-p/1712

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluations of teaching: The state of the art. *Review of Educational Research, 83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Uttl, B., White, C. A., & Wong Gonzalez, D. (2017, September). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54,* 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wieman, C. (2015). A better way to evaluate undergraduate teaching. *Change: The Magazine of Higher Learning, 47*(1), 6–15. https://doi.org/10.1080/00091383.2015.996077