Teaching Quality Framework (TQF)

Frequently Asked Questions

Draft: May 31, 2022

Please Note: This document is currently in development. *It is intended to provide guidance and resources to those interested in improving teaching assessment in higher education. The recommendations and philosophy of assessment described below will continue to evolve with additional resource-gathering and research.*

This document's development has been supported by funding from the National Science Foundation (DUE-1725959), the American Association of Universities, and the University of Colorado Boulder. Any opinions, conclusions, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of those organizations.

For more information about the Teaching Quality Framework, please visit our website:
https://www.colorado.edu/teaching-quality-framework/

Teaching Evaluations (General Issues and TQF Philosophy)

Why change our current teaching evaluation practices?

1. At many research universities, current evaluation systems for tenure and promotion poorly assess and/or undervalue the importance of teaching effectiveness, relying instead on research success (Bradforth et al., 2015). Current practices at most research universities have led to limited value being placed on effective teaching at best, and — at worst — actually reward poor teaching (Braga, Paccagnella, & Pellizzari; 2014).
2. Normative practices are neither as effective as we want nor do they give pathways for improvement; that is, they often lack systematic processes for formative development es and can provide inaccurate summative information. For more information on this, see the section on SETs.
3. There is growing interest at the national level (PCAST, 2012; NASEM, 2018) to improve the quality of teaching. Research on institutional change, effective teaching practices, and the scholarship of teaching and learning now enable us to develop a research-based evaluation framework. For a more thorough justification, see the TQF white paper (Finkelstein et al., 2017).

What is the role of formative and summative assessment in an evaluation system?

1. There is a need to balance teaching growth of all instructional faculty (formative) against the need to evaluate instructors for reappointment, tenure and/or promotion (summative). In order to achieve this balance we recommend that evaluation processes include both formative and summative assessment. Summative assessment results in a determination of whether an instructor is effective or not, while formative assessment results in recommendations for improving teaching. Summative evaluation results in a product (a grade of teaching effectiveness), while formative evaluation continues a process (the refinement of teaching).
2. A recent study (Lyde, Grieshaber, & Byrns, 2016) that interviewed faculty on an evaluation system involving multiple methods of evaluation (student evaluations of teaching, faculty portfolios, and self-reflection) found that faculty felt that the evaluations were disproportionately summative, lacking sufficient formative outcomes. This study provides a good example of how formative evaluation can be used to refine the evaluation systems of faculty.

Why three voices for assessment (peer, self, student)?

1. **Peer voice** allows for instructor evaluations to be more holistic, as instructors are qualified to evaluate teaching elements such as the instructor's content understanding and course goals. Additionally, peer observers who are involved in the education research community are qualified to evaluate if the instructor is using up-to-date instructional strategies. See *Role of Peers in Evaluation: Peer Voice* for more information on peer voice.
2. **Faculty themselve**s are uniquely able to describe their justifications for teaching goals and course design, their efforts for professional development, the learning outcomes of their students, and their engagement with the scholarship of teaching. See *Role of Faculty/Instructors in Their Own Evaluation: Faculty Voice* for more information on faculty voice.
3. **Students** can provide feedback on areas such as engagement, clarity, and their satisfaction with the course. They can provide comparisons between courses on difficulty, time required for the

course, as well as information about themselves such as their perceived efforts relative to other students or confidence with the content before taking the course. Lastly, they can demonstrate awareness of learning goals and objectives (e.g. critical thinking skills, mastery of the material, ability to work on a team, and engage in independent learning, etc.). See *Role of Students in Evaluation: Student Voice (incl. SETs)* for more information on student voice.

Why is it important to use multiple measures?

1. There are some questions that cannot be adequately answered by only one of the three voices (students, peers, or the faculty member being evaluated) or individual measures (e.g., student ratings, faculty teaching practices inventories vs. teaching portfolios). Just as students are ill-equipped to judge a faculty's understanding of content or course design, peer evaluators are unable to effectively comment on student interactions with the instructors. By using multiple measures, particularly when they include the three voices, we can effectively "span the space" of information that can be gathered about an instructor's teaching effectiveness.
2. Bias is inherent in any evaluation, and one of the best ways to mitigate bias is to have multiple sources for the evaluation. For more information on this, see the Evaluation Bias section of the FAQ.
3. Most institutions require multiple measures. For example, the University of Colorado requires a minimum of three measures of faculty evaluation (https://www.cu.edu/ope/aps/1009). As a note, Colorado's only required measure is using student evaluations of teaching.

Why does TQF emphasize departmental engagement?

1. The department has been identified as the key unit of sustained change in research-extensive universities, like the University of Colorado (CU) (AAU et al., 2017).
2. Our approach adopts the Departmental Action Team (DAT) model for engaging departments (Corbo, 2015). DATs are working groups with 4-6 faculty and 1-2 facilitators, designed to involve members of the department in the development and implementation of change efforts to make sustainable reform.
3. While TQF is attentive to the department as a key unit of change, it is not the only location of work. A systems view of teaching evaluation at the university will involve attention to individual faculty, professional development programs, college and administrative levels, as well as discipline-based professional societies and extramural stakeholders. For example, recent efforts at discipline-based societies (e.g., AAPT), professional organizations (e.g., AAU, 2015), accreditors (e.g., HLC, 2016), and legislative bodies (e.g., CO Senate Bill 10-191, relating to K-12 teaching assessment) represent a broader trend in attention to teaching evaluation by extramural stakeholders.

Evaluation Systems

What are better practices for incorporating multiple measures into an evaluation system?

1. Know what your goals are and what tools meet those goals. Some questions on end-of-term student ratings (SETs/FCQs) adequately measure student satisfaction, perception of workload,

etc. Other tools, such as observations of classroom practices (e.g., COPUS), are helpful for understanding how classroom time is spent and what teaching practices are employed.

2. Consider the use of a rubric for looking at the multiple dimensions and ensuring that the measures provide a holistic view of the instructor. The Teaching Quality Framework Assessment Rubric and other resources are available here.
3. Use existing services and tools rather than making them up. There is extensive literature on using measures for teaching evaluation (e.g., Evaluation of teaching: Challenges and promises). In addition, existing tools can be adapted to suit local environments. Many units at CU Boulder have worked with the TQF to develop tools and processes for teaching evaluations; a repository of publicly available examples can be found here. Additional institutional teaching evaluation resources can be found here. Please contact us if you are looking for a particular resource that cannot be found on our site. Additional CU Boulder resources include:
    a. Center for Teaching and Learning (CTL) Programs & Services
    b. ASSETT (A&S) services for class observations and more.
4. Have faculty engage in self-reflection (reviewing their own practices) in a structured and systematic matter.
5. When using materials, focus on the engagement of faculty in using measures to assess their own practices with an emphasis on formative assessment.

Are there evaluation tools that distinguish between good and exceptional instructors?

1. Weisberg et al. (2009) found that over 90% of instructors receive the top two ratings on a point scale for evaluation, with less than 1% receiving an unsatisfactory mark. When instructors were asked to rate themselves on a scale of 1 to 10, over 85% rated themselves an 8 or above. This suggests that individual tools may not be able to distinguish between good and exceptional instructors, which is one reason it is important to use multiple measures of teaching effectiveness across multiple voices. The same report suggests developing a comprehensive evaluation plan, training administration in conducting evaluation, and rewarding good teaching.
2. One evaluation system that has found success is TAP. From the National Institute for Excellence in Teaching, Daley & Kim (2010) report on the TAP system that was able to produce a normal distribution of instructor evaluation ratings, differentiating good instructors from exceptional instructors. A key point for their success is the use of measurement across multiple (4) dimensions with multiple ratings for each dimension that are averaged out.
3. As an alternative to "grading" instructors on a fixed-point scale, instructors can be promoted to higher roles reflecting their cumulative capabilities as an instructor. An example of this is the Royal Academy of Engineering Career Framework for University Teaching (RAENG, 2018).

How can evaluation systems be used to value and incentive effective teaching?

1. Promotion and merit decisions at most research universities rely on instructors reaching a set level of aggregate contributions through research, teaching, and service. This is described as the one-bucket system, where all three factors contribute to filling up a bucket, where exemplary research can balance out poor instruction, or vice-versa. However, often the level required for promotion can be reached solely on research efforts, disincentivizing teaching and service. To value teaching and service, the University of California (UC) Irvine has begun transitioning to a

three-bucket system. In the three-bucket system, research, teaching, and service each have their own buckets and levels must be reached for each factors. For detailed descriptions on how UC Irvine uses the three-bucket system for promotion and merit decisions, see the *Criteria for Appointment, Promotion, and Appraisal* section of their Appointment and Promotion guidelines (https://www.ucop.edu/academic-personnel-programs/_files/apm/apm-210.pdf).

2. Departmental definition of assessment measures and relative weights of research, teaching, and service is essential in defining faculty expectations and provides a roadmap for faculty members' continuous improvement of their teaching.
3. The traditional approach to evaluating teaching examines only the behavior of the instructor in the classroom and uses limited evidence, typically only student evaluations of teaching. The University of Kansas (KU) developed a rubric to measure 7 dimensions in teaching to provide a holistic view of the instructor as an educator. The TQF has a modified version of this rubric available on the resources page of the TQF website.
4. For more information on rewarding teaching, see Dennin et al. (2017).

What are things to consider when establishing an evaluation system?

1. Ensure that the evaluation measures lead to actionable outcomes. The measures should be reliable (repeatable across different settings), valid (measures should judge instructors in a reasonable way, like the *don't judge a fish on its ability to climb a tree* adage), fair (the measures minimize bias), and avoid undesirable consequences of using the measure (e.g. instructors making a class easier due to the discredited belief that students rate easy classes better).
2. Fairness and reliability are most easily addressed by using multiple measures. See the Evaluation Bias section for more information.
3. Disciplinary action and approaches for remediation of unsatisfactory evaluation results should be defined.
4. While drawing from evidence-based tools, be sure to consider how these tools are used locally, which are the best-suited tools and approaches to a given context, and how they should be adapted to fit local needs.
5. While adaptation is encouraged, there should be enough commonality across the different units that measures used in the evaluation system are coherent across the campus.
6. It can be easier to establish a new approach to evaluation by starting with low-stakes improvements, with evaluations used for merit rather than reappointment, promotion, and tenure. Additionally, value participation in the new evaluation system; that is, reward faculty for having trust and being willing to engage in new evaluation measures.

How can evaluation best be used to improve teaching? [in development]

1. When considering formative feedback, comments should be specific and focus on behaviors and practices rather than the instructor. Comments should be also be provided to the instructor soon after the evaluation measurement (Brinko, 1993).
2. Expectations for instruction should be explicit and clear to instructors. This offers several benefits, including guiding expectations for faculty publicly and framing resources for improvement, such as professional development programs.

Evaluation Bias

What bias is present in evaluations?

1. There are a few different types of bias. The easiest to control is bias in administration procedures. This bias is seen most easily in SETs. For example, de-anonymizing SETs or having the instructor present during administration of the SETs both lead to artificial inflation in ratings. For example, Youmans and Jee (2007) found that offering chocolate to students led to an increase in evaluation scores. To best control this form of bias, departments should create a standardized method of administering SETs, and in conducting instructor evaluations in general.
2. Another form of bias is implicit bias, for example gender and race bias. Regarding gender bias, there is little agreement in results. Studies that have falsified instructor gender or created fictitious instructor profiles have found significant differences in ratings based on gender (MacNell, Driscoll, & Hunt; 2015), while others that have studied real SETs have found no significant difference (Feldman, 1993). The discrepancies in researchers' conclusions may be due to variation in bias across fields/disciplines and other contextual factors.
3. Bias exists in SETs between disciplines, with ratings significantly lower for STEM/quantitative disciplines than non-STEM/non-quantitative disciplines. For a review of the effects of instructor gender and discipline on student ratings, see IDEA Research Report #10 (Li & Benton 2017).

How do we address bias in the system? [in development]

1. Raise awareness of implicit bias
2. Make framework/criteria for evaluation explicit and public
3. Use multiple measures of teaching effectiveness.

Leveraging Three Voices of Teaching Assessment (Peer, Self, Student)

Role of Peers in Evaluation: Peer Voice

How might we incorporate peer/classroom observation? [in development]

1. Peer observation is the most common method of engaging peer voice in evaluation. There are many well-researched observation protocols available; many of these have been adapted by units here at CU Boulder. Examples can be found here. Please contact us if you are looking for a particular resource that cannot be found on our site.
2. The University of Colorado, through the Arts & Sciences Support of Education Through Technology (ASSETT) program, offers the Visualizing Instructional Practices (VIP) service. Upon request, ASSETT staff will conduct a classroom observation using the Classroom Observation Protocol for Undergraduate STEM (COPUS).

What are other ways to incorporate peer voice or peer review into assessment? [in development]

1. Course Portfolios / Review (Bernstein, Burnett, Goodburn, & Savory, 2006)
2. [In development...others to be added]

Role of Faculty/Instructors in Their Own Evaluation: Faculty Voice

What are the suggested forms of faculty voice?

1. There are several avenues for faculty to reflect on their instruction.
    a. A statement of teaching philosophy is a short narrative describing the instructor's thoughts on teaching and learning, how they teach, and why they teach that way.
    b. Self-reflection inventories are questionnaires designed to evaluate an instructor's teaching practices, such as questions about the use of research-based instructional strategies.
        i. An extension of self-reflection inventories are self-efficacy inventories. Rather than probe at teaching practices, efficacy inventories aim to provide understanding of an instructor's belief that they can succeed at teaching.
    c. Teaching portfolios are comprehensive dossiers that provide a holistic view of the instructor. These typically include course materials such as syllabi, assignments, and exams, as well as evaluations of the instructor, whether they be from students, peer observations, and the instructor's self-reflections and statement of teaching philosophy. The portfolio should give readers insight into the instructor's achievements as well as the instructor's approach to teaching.
    d. Instructors can maintain reflection journals, where they write narratives directly after each class. Instructors can discuss things such as what went well and what should be changed in the future.
    e. Instructors can record themselves teaching for later review, potentially using observation protocols. Using this method, instructors can approach self-reflection the same way that peer observation is approached.
    f. Many units at CU Boulder have worked with the TQF to develop tools and processes for self-evaluation of teaching. Publicly available examples can be found here. Please contact us if you are looking for a particular resource that cannot be found on our site.

Role of Students in Evaluation: Student Voice (incl. SETs)

Can SETs be effective at measuring learning outcomes?

1. A recent meta-analysis by Uttl et al. (2017) found that there is little to no correlation between student learning gains and SETs. The study concludes that students are not likely to learn more from an instructor with better SET ratings. They follow with claims as to why this is the case, citing that student's interest, motivation, prior knowledge, and other factors lead to SETs being unreliable. Another study by Uttl et al. (2013) found that SETs were nearly six standard deviations lower for quantitative courses than non-quantitative courses.
2. A common issue with past SETs is that students are asked to rate instructors on areas they are ill-equipped to judge. For more information, see the Evaluation Systems section of this FAQ.

What makes for better SET questions?

1. One well-researched student evaluation form is the Teaching Behavior Checklist (TBC) (Keeley et al., 2009). A factor analysis of the TBC found that the qualities measured can be grouped into two distinct categories: the instructor's professional competency/communication skills, and caring/support provided to students. Similarly, another resource is the Student Assessment of Learning Gains (SALG).
2. Questions should be focused on the student perspective, such as their interactions with faculty, their awareness of course goals, and the demands on them in and out of class.
3. Questions can prompt narrative responses rather than multiple choice, Likert-scale style questions. An example of this approach can be seen in Yale's end-of-term course evaluation system (Yale Recommendations to Revise the Yale College Online Course Evaluation Survey, Report of the Teaching and Learning Committee 2015-16; Current Policies).

Can online SETs match the response rate of in-person SETs?

1. Historically, online response rates have been significantly lower than in-class response rates. The gap in response rates decreased as class size increases, but at best online response rate is 15% worse than in-class. For a large-N review, see IDEA Research Notes #4.
2. Nissen et al. (2018) found that online response rate could match in-class response rate for research based assessments when sufficiently motivated. In practice, to get high response rates for online administration of SETs, instructors should: send multiple email reminders, give multiple in-class reminders, and provide participation credit for completion. It is important to note that online response rates were lower (though not statistically significantly lower) for students who had lower course grades as compared to in-class response rates.

How can student voice be used beyond end-of-term evaluation forms/SETs?

1. Mid-semester feedback (MSF) can be helpful in engaging student voice for instructor evaluation. Historically, MSF has focused entirely on student experience, and thus has not been plagued with many of the reliability and validity issues present in SETs. Rather than rate faculty effectiveness, in MSF, students have the opportunity to discuss what has and hasn't worked for them so far and provides faculty with the opportunity to more readily see the effects of making adjustments. For a review of the benefits of MSF and a thorough guide on effectuating MSF, see the Midterm Student Feedback Guide (http://bit.ly/msfguidebook).
    a. CU offers MSF in the form of the Classroom Learning Interview Process (CLIP). CLIPs are focus-group style interviews with students during usual class time, conducted by peer faculty.
    b. For an example of a written, easy-to-use, single-page MSF form, see the Stop-Go-Change Evaluations (Sayre, accessed December 2018), available at PhysPort.

Will students give higher ratings for easier classes?

1. A common concern is that students will rate classes poorly if they are challenging, and rate classes higher if they think they'll do better. Marsh (1987) found that a majority of instructors believe that difficulty, workload, and stringency of grading all negatively bias student ratings. Stroebe (2016) argues that the valuing of SETs has led to grade inflation based on studies that

show expected grades correlate with student ratings, examined using manipulated grades. However, when real grades and student grade expectations are compared, Centra (2003) found that expected grades don't affect SETs. Regarding course difficulty, Centra found there is a middle ground for class difficulty. Classes that were not only too difficult, but also ones that were too easy, received lower ratings than classes that found an appropriate balance.

Teaching practices

What are key teaching behaviors that are associated with learning outcomes?

1. One of the most impactful changes that can be made in the classroom is engaging students. Interactive engagement has shown to be far more effective regarding student understanding of content than traditional lecturing (Hake, 1998). Additionally, failure rates are lowered and student success is increased in active learning environments as compared to traditional lecture (Freeman, 2014).
2. Engaging in non-traditional classroom practices such as collaborative learning and conceptual problem solving led to increased student self-efficacy (Fencl & Scheel, 2005).
3. For STEM courses, the National Academies' Consensus Study Report on STEM Education (Reaching Students, 2015) and the White House's PCAST Report (Engage to Excel, 2012) have described a variety of interactive practices and approaches that support enhanced student learning.

Establishing Learning Goals

Why should instructors identify learning goals?

1. Identifying learning goals and sharing them with the students is useful for course development and contributes to department-wide knowledge of what students will learn after taking a course. Additionally, students often rely on them and find them helpful once learning goals becomes a recurring feature in courses.
2. Although there are many potential learning goals, some will be more important to the instructor and the purpose of the course than others. For example, learning to think ethically may be more valued in a journalism course than an intro-level physics lab, whereas the opposite may be true for the goal of improving quantitative literacy. By focusing in on key learning goals, the instructor can use teaching methods that are shown to be particularly effective at attaining those goals. For a list of teaching methods associated with learning objectives, see IDEA Technical Report #19 (Li et al., 2016), with a summary figure here.

How can instructors identify learning goals?

1. The Science Education Initiative (SEI) Handbook (Chasteen & Code, 2018) provides a guide of questions to ask when developing learning goals, and addresses concerns such as whether learning goals lead to "teaching to the test" or if they restrict what an instructor can teach.

2. A strong baseline to consider while developing learning goals is the American Association of Colleges and Universities (AACU) Liberal Education and America's Promise (LEAP) project which identify several essential learning outcomes.
    a. Content understanding, with a focus on big questions
    b. Intellectual and practical skills, such as critical thinking and communication skills
    c. Personal and social responsibility, including ethical reasoning and establishing the foundations for lifelong learning
    d. Applied learning, including the ability to synthesize knowledge across domains

References

1. Association of American Universities. (2015). Aligning Practices to Policies. Washington, DC. https://www.aau.edu/sites/default/files/AAU-Files/STEM-Education-Initiative/Aligning-Practice-To-Policies-Digital.pdf
2. Association of American Universities. (2017). Progress Toward Achieving Systemic Change: A Five-Year Status Report on the AAU Undergraduate STEM Education Initiative. Washington, DC. https://www.aau.edu/sites/default/files/AAU-Files/STEM-Education-Initiative/STEM-Status-Report.pdf
3. Benton, S. L., & Cashin, W. E. (2013). Student Ratings of Teaching: A Summary of Research and Literature. IDEA Paper #50. Manhattan, KS: The IDEA Center.
4. Bernstein, D. J. (2008). Peer Review and Evaluation of the Intellectual Work of Teaching. Change: The Magazine of Higher Learning, 40(2), 48-51. https://doi.org/10.3200/CHNG.40.2.48-51
5. Bernstein, D., Burnett, A. N., Goodburn, A., & Savory, P. (2006). Making Teaching and Learning Visible: Course Portfolios and the Peer Review of Teaching. https://www.wiley.com/en-us/Making+Teaching+and+Learning+Visible%3A+Course+Portfolios+and+the+Peer+Review+of+Teaching-p-9781882982967
6. Bradforth, S. E., Miller, E. R., Dichtel, W. R., Leibovich, A. K., Feig, A. L., Martin, J. D., Bjorkman, K. S., Schultz, Z. D., & Smith, T. L. (2015). University learning: Improve undergraduate science education. Nature, 532(7560), 282-284. https://doi.org/10.1038/523282a
7. Braga M., Paccagnella M., & Pellizzari M. (2014). Evaluating students' evaluations of professors. Economics of Education Review, 41, 71-88. https://doi.org/10.1016/j.econedurev.2014.04.002
8. Brinko, K. T. (1993). The Practice of Giving Feedback to Improve Teaching: What Is Effective? The Journal of Higher Education, 64(5), 574-593. https://dx.doi.org/10.2307/2959994
9. Centra, J. A. (2003). Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? Research in Higher Education, 44(5), 495-518. https://doi.org/10.1023/A:1025492407752
10. Chasteen, S. V., & Code, W. J. (2018). The Science Education Initiative Handbook. https://pressbooks.bccampus.ca/seihandbook/
11. Corbo, J., Reinholz, D. L., Dancy, M. H., & Finkelstein, N. F. (2015). Departmental Action Teams: Empowering faculty to make sustainable change. Paper presented at Physics Education Research Conference 2015, College Park, MD. https://doi.org/10.1119/perc.2015.pr.018
12. Daley, G., & Kim, L. (2010). A teacher evaluation system that works. NIET Working Paper. Santa Monica, CA: The National Institute for Excellence in Teaching. https://files.eric.ed.gov/fulltext/ED533380.pdf
13. Dennin, M., Schultz, Z. D., Feig, A., Finkelstein, N., Greenhoot, A. F., Hildreth, M., Leibovich, A. K., Martin, J. D., Moldwin, M. B., O'Dowd, D. K., Posey, L. A., Smith, T. L., & Miller, E. R. (2017). Aligning Practice to Policies: Changing the Culture to Recognize and Reward Teaching at Research Universities. CBE Life Sciences Education 16(4). https://doi.org/10.1187/cbe.17-02-0032
14. Dunn, K. E., & Mulvenon, S. W. (2009). A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education. Practical Assessment, Research & Evaluation, 14(7). https://pareonline.net/pdf/v14n7.pdf

15. 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, 34(2), 151-211. https://doi.org/10.1007/BF00992161
16. Fencl, H. & Scheel, K. (2005). Engaging Students: An Examination of the Effects of Teaching Strategies on Self-Efficacy and Course Climate in a Nonmajors Physics Course. Journal of College Science Teaching, 35(1), 20-24. https://www.jstor.org/stable/42992548
17. Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Nnadozie, O., Jordt, H., and Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. PNAS 111(23), 8410-8415. https://doi.org/10.1073/pnas.1319030111
18. Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics 66(1), 64-74. https://doi.org/10.1119/1.18809
19. Higher Learning Commission. (2016). Determining Qualified Faculty Through HLC's Criteria for Accreditation and Assumed Practices. Chicago, IL. http://download.hlcommission.org/FacultyGuidelines_2016_OPB.pdf
20. IDEA. (2010). Research Notes #4: Paper versus Online Survey Delivery. Manhattan, KS: The IDEA Center. https://www.ideaedu.org/Portals/0/Uploads/Documents/Research%20Notes/IDEA_Research_Notes_4-Online_v_Paper_Delivery.pdf
21. Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating Psychology Students' Perceptions of Teachers Using the Teacher Behavior Checklist. Teaching of Psychology 37(1), 16-20. https://doi.org/10.1080/00986280903426282
22. Li, D., Benton, S. L., Brown, R., Sullivan, P., & Ryalls, K. R. (2016). Analysis of IDEA Student Ratings of Instruction System 2015 Pilot Data. IDEA Technical Report #19. Manhattan, KS: The IDEA Center. http://www.ideaedu.org/Portals/0/Uploads/Documents/Technical-Reports/IDEA_Technical_Report_No_19.pdf
23. Li, D., & Benton, S. L. (2017). The Effects of Instructor Gender and Discipline Group on Student Ratings of Instruction. IDEA Research Report #10. Manhattan, KS: The IDEA Center. https://www.ideaedu.org/Portals/0/Uploads/Documents/Research%20Reports/Research_Report_10.pdf
24. Lyde, A., Grieshaber, D., & Byrns, G. (2016). Faculty Teaching Performance: Perceptions of a Multi-Source Method for Evaluation. Journal of the Scholarship of Teaching and Learning, 16(3), 82-94. https://doi.org/10.14434/josotl.v16i3.18145
25. MacNell, L, Driscoll, A., & Hunt, A. N. (2015). 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
26. Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Education Research, 11(3), 253-288. https://doi.org/10.1016/0883-0355(87)90001-2
27. National Academies of Sciences, Engineering, and Medicine (2018). Indicators for Monitoring Undergraduate STEM Education. Washington, DC: The National Academies Press. https://doi.org/10.17226/24943.

28. National Research Council (2015). Reaching Students: What Research Says About Effective Instruction in Undergraduate Science and Engineering. Washington, DC: The National Academies Press. https://doi.org/10.17226/18687
29. Nissen, J. M., Jariwala, M., Close, E. W., & Van Dusen, B. (2018). Participation and performance on paper- and computer-based low-stakes assessments. International Journal of STEM Education, 5(1), 21. https://doi.org/10.1186/s40594-018-0117-4
30. President's Council of Advisors on Science and Technology. (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President. Executive Office of the President. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf
31. Royal Academy of Engineering (2018). Career Framework for University Teaching: background and overview. http://www.raeng.org.uk/CareerFrameworkUniversityTeaching
32. Sayre, E. Stop-Go-Change: Midterm Evaluations. Manhattan, KS. https://www.physport.org/recommendations/files/Stop-Go-Change-Evals.pdf
33. Carroll, S., Seymour, E., and Weston, T. (2007). Student Assessment of their Learning Gains. https://salgsite.net/
34. Stroebe, W. (2016). Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations. Perspectives on Psychological Science 2016, 11(6), 800–816. https://doi.org/10.1177/1745691616650284
35. University of Kansas Center for Teaching Excellence (2016). Department Evaluation of Faculty Teaching Rubric. Lawrence, KS. https://cte.ku.edu/benchmarks-teaching-effectiveness-project
36. Uttl, B., White, C. A., & Morin, A. (2013). The Numbers Tell It All: Students Don't Like Numbers! PLoS ONE 8(12): e83443. https://doi.org/10.1371/journal.pone.0083443
37. Uttl, B., White, C. A., Wong Gonzalez, D. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation, 54, 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007
38. Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New York: The New Teacher Project. https://files.eric.ed.gov/fulltext/ED515656.pdf
39. Yale College Teaching & Learning Committee (2016). Report of the Teaching and Learning Committee 2015-16. New Haven, CT. https://yalecollege.yale.edu/sites/default/files/files/OCE%20Committee%20Report%2009-2016.pdf
40. Yale Poorvu Center for Teaching and Learning. End-of-Term Evaluations. New Haven, CT. https://poorvucenter.yale.edu/End-of-Term-Evals
41. Youmans, R. J., & Jee, B. D. (2007). Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course. https://doi.org/10.1080/00986280701700318