Preliminary takeaways from the IR/TQF FCQ pilot project at CU Boulder*

Fall 2020

As we consider how we might deploy and interpret FCQs as we adapt to the COVID-induced changes, we share takeaways from an effort to implement better questions and processes for interpreting results from the FCQs.

Why do universities want better student evaluations of teaching (SETs, or FCQs at CU)?

- Many SETs ask questions students are ill-equipped to evaluate (e.g., how much they learned)[1]
- SETs are rarely correlated with learning gains[2]
- Bias:
    - biased ratings can occur for women/faculty of color, particularly in STEM[3,4]
    - quantitative classes yield lower scores[4]
    - low response rates for online SETs can bias outcomes[5]
- Poorly-designed SETs may incentivize unproductive teaching practices (e.g., grade inflation)[6]
- Most current FCQ questions at CU Boulder do not provide specific actions for improvement

Some takeaways from preliminary analyses of three semesters of pilot FCQ data

(see here for the 2019-10-04 TQF stakeholders' meeting presentation; below is a summary)

1. What you see is what you get
    a. There is some statistically-significant evidence of systematic bias, but effect sizes are very, very small
    b. Patterns of responses observed in simple descriptive (bivariate) analyses are generally born out in more complex (multilevel, multivariate) analyses, so basic visualizations will generally "tell the story"
    c. Modeling predictors for the pilot items and standard items yield similar patterns. Scores on current FCQ items and pilot items are highly correlated and yield similar patterns of predictors
2. Use FCQs to flag potential concerns
    a. FCQs (standard and pilot) do not distinguish good instructors from great instructors and barely differentiate average instructors from great instructors, but they can help flag concerns
    b. Small differences likely don't mean much, especially at average or above-average FCQ scores; differences of one point or more (on the six point scale) may signal a difference.
3. Choose questions for non-measurement qualities
    a. Choose FCQ questions that are important (e.g., inclusivity), provide actionable feedback, and tap into students' experiences or observations of practice
4. Use multiple measures
    a. FCQs should not be standalone measures of teaching quality
    b. Examples: peer observation, self-reflection, other forms of student voice such as classroom interviews, etc.
5. Test other improvements
    a. Continue to develop better student evaluations of teaching (e.g. improving question sets, better visualizations, useful comparisons)
    b. For example, some FCQ challenges *may* be mitigated by improved instructions and training for students, and changing norms around FCQ responding to yield higher quality responses

While there is little difference in the measurement qualities of the pilot and standard questions, the pilot questions may be better than the standard questions in that they provide more actionable feedback and tap into students' experiences/observations of practice.

* Based on pilot studies conducted by the Office of Data Analytics (ODA), and analysis conducted by ODA and the TQF. Prepared by the Teaching Quality Framework (TQF) Initiative (http://www.colorado.edu/teaching-quality-framework) with input from Jess Keating.

The 8 core pilot IR/TQF questions

In this course, I was encouraged to:
1. Reflect on what I was learning or how I was learning.
2. Evaluate arguments, evidence, assumptions, and conclusions about key concepts (critical thinking).

In this course, the instructor:
3. Maintained an environment that was respectful of diverse students and diverse points of view.
4. Seemed personally invested in student success.
5. Provided content and materials that were helpful.

In this course, I was:
6. Challenged to develop my own knowledge, comprehension, and conceptual understanding.
7. Provided opportunities to ask questions and initiate discussion.
8. Provided feedback on my work that helped me improve my performance.

Note in In spring 2020, due to emergency shifts related to the COVID-19 pandemic, CU Boulder adopted a new FCQ question set that was developed by the FCQ Redesign Project (2016-2017); in October 2020, the Provost announced that this question set would continue to be used in fall 2020 and thereafter. Most of the IR/TQF pilot items above are included in this new question set. A comparison of the new 2020 FCQ question set to the IR/TQF pilot items can be found here: https://www.colorado.edu/teaching-quality-framework/IR-TQF_FCQ_Pilot_v_2020_FCQ

References

[1]Kunter, M., & Baumert, J. 2006. Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3): 231–251. http://dx.doi.org/10.1007/s10984-006-9015-7

[2]Uttl, B., White, C.A., & Gonzalez, D.W. 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

[3]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

[4]Smith, B.P., & Hawkins, B. 2011. Examining student evaluations of Black college faculty: Does race matter?. *The Journal of Negro Education*, 80(2): 149-162. https://www.jstor.org/stable/41341117

[5]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

[6]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

[6]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*, 11(6): 800–816. https://doi.org/10.1177/1745691616650284