

# INTERPRETING STUDENT EVALUATIONS OF TEACHING

Victoria Akin<sup>1</sup> and Shira Viel<sup>1</sup>

**Importance.** Student Evaluations of Teaching (SET) are a ubiquitous tool in the assessment of instructor effectiveness. SET are prevalent because they are simple to administer, and results are assumed to be easy to present and interpret [9]. Taken together with classroom observations and other assessment methods, SET can provide productive feedback to college instructors. However, one must be judicious in the use of these instruments and avoid using them as the only method of evaluating teaching performance. For graduate teaching assistants (GTAs) to benefit from SET, they need to know the strengths and weaknesses of this type of assessment tool. To emphasize the need for care in analysis and interpretation of the data they provide, SET are often referred to as student ratings rather than evaluations [1, 5]. This summary briefly discusses the utility and limitations of student ratings in terms of their validity and reliability, and concludes with recommendations for their effective use. Note that we primarily cite summary studies: the interested reader should consult the references therein.

**Validity.** The *validity* of an assessment tool is the extent to which the tool accurately measures the quality or quantity intended to be measured. Student ratings are intended to measure “quality of teaching.” However, there is no consensus definition of effective teaching, making appraisal and communication of results difficult. Further, there are many factors unrelated to teaching effectiveness that bias student ratings. A survey paper by Kornell and Hausmann [7] includes the following list of items that may impact student ratings: “student gender, prior subject interest, and expectations for the course; instructor gender, ethnicity, attractiveness, charisma, rank, and experience; and course subject area, level, and workload” (p. 2). In particular, many studies suggest that gender negatively affects student ratings of female teachers, even in online courses [3]; the results of MacNell, Driscoll, and Hunt [8] are exceptionally compelling. Despite these biases, student ratings have been shown to correlate with outcomes commonly associated with effective teaching. In some studies, higher numerical scores on teaching evaluations correlate positively with higher scores on common exams [1, 4]. However, the instructor qualities that promote achievement in one course could decrease the deep learning of transferable knowledge that would assist performance in subsequent classes [4, 7]. Overall, reviews of the current literature present an inconclusive picture of SET validity [see, for example, 1, 9], making them suspect as a tool for gauging the quality of an instructor's teaching.

**Reliability.** The *reliability* of an assessment tool refers both to stability, the agreement between ratings given at different points in time, as well as consistency, the agreement between ratings given at a specific point in time. It is important to emphasize that a reliable assessment need not be valid, and some argue that the reliability of student ratings is irrelevant due to their invalidity [10]. In general, instructor ratings tend to be relatively stable over time [1], although a recent working paper found a decline following tenure [6]. Consistency within a given term tends to increase with the number of raters [1]. Another facet of reliability is *generalizability*, the extent to which student ratings capture overall teaching effectiveness irrespective of the course or term. While some studies conclude that the instructor, rather than the course or term, is the primary factor determining student ratings [1], analyzing results from a variety of courses over a variety of terms can increase the generalizability of the data.

**Use.** Research indicates that getting feedback from midterm student ratings is positively associated with receiving higher end-of-term student ratings, and that this positive effect is increased when the feedback is combined with consultation [1]. Due to the biases inherent in student ratings, it is recommended that they be used primarily for the formative purpose of making adjustments in individual instruction, and only be used for summative purposes, such as promotion and reappointment, in conjunction with other assessments of teacher quality [2, 5].

## Example Professional Development Resources:

### Example Resource 1. Video Cases for College Math Instruction: Processing Student Feedback

Location: <http://collegemathvideocases.org/cases/case.php?VCID=4>

**Goals:** This activity aims to prepare GTAs for reading, responding to, and acting on end-of-term SET. The first part of the activity uses video clips of two GTAs' classroom instruction to assist participants in identifying and distinguishing between classroom characteristics over which the instructor has control and those for which modification may be more limited. The second part of the activity uses video clips of the same two GTAs reading end-of-term SET aloud to help participants practice synthesizing feedback from individual students and deciding if and how any common instructional issues should be addressed in future classrooms.

**Synopsis:** The activity has two parts, each built around short online video clips of GTAs from another institution. It is designed to be completed with a group of GTAs at any point during their GTA professional development. Each part of the activity has a similar structure: first participants review the discussion questions, then watch the videos while taking

---

<sup>1</sup>Duke University

## INTERPRETING STUDENT EVALUATIONS OF TEACHING

Victoria Akin and Shira Viel

notes (for the first part, the clips are watched twice), and conclude by processing the discussion questions. While the first part, built around clips of the GTAs' instruction, need not have a full-group discussion afterwards, the second part is designed as a consensus-building exercise. After watching the GTAs read their SET, participants are asked to arrive at a single concrete list of recommendations for instructional change.

### Example Resource 2. Self Evaluation Exercise

Location: Files are in the College Mathematics Instructor Development Source (CoMInDS) community at MAA Connect [connect.maa.org](http://connect.maa.org)

Goals: This activity encourages first-time GTAs to read carefully and productively through student responses on their end-of-term SET. The activity provides a framework for interpreting student comments and ratings in a constructive way. While reflecting on their own teaching, GTAs should identify areas for improvement and make plans for incremental change. The activity aims to help GTAs avoid feeling demoralized by negative criticism and instead stay motivated to teach effectively in the future. By including a meeting with an experienced instructor, the activity is designed to help novice instructors maintain a broader perspective and recognize that there are many aspects of effective teaching that are not necessarily captured by SET. The self-evaluation exercise puts new instructors in contact with other members of their community who are invested in student learning, and helps build a mentoring relationship between new and experienced instructors.

Synopsis: The activity has four parts. It is designed to be completed by an individual GTA at the end of their first term teaching, in conjunction with an experienced instructor. First, the GTA reads through their SET after the end of their first course. The GTA then reflects on their first term teaching and writes a response to the student comments. Next, the GTA sends the written response to a GTA professional development provider. The GTA and experienced instructor then meet to discuss both the SET and the reflection.

### References

1. Benton, S. L., & Cashin, W. E. (2011). IDEA Paper No. 50: Student ratings of teaching: A summary of research and literature. Manhattan, KS: The IDEA Center.  
[http://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA\\_50.pdf](http://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_50.pdf)
2. Benton, S. L., & Ryalls, K. R. (2016). IDEA Paper #58: Challenging misconceptions about student ratings of instruction. Manhattan, KS: The IDEA Center.  
[http://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA\\_58.pdf](http://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_58.pdf)
3. Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research 2016* (DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1).
4. 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.
5. Dewar, J. (2017). Student ratings of teaching: What every instructor should know. *AMS Blogs*, April 17, 2017. <https://blogs.ams.org/matheducation/2017/04/17/student-evaluations-ratings-of-teaching-what-every-instructor-should-know/>
6. Gourley, P., & Madonia, G. (2019). Tenure and Faculty Course Evaluations.  
[https://www.researchgate.net/publication/331982853\\_Tenure\\_and\\_Faculty\\_Course\\_Evaluations](https://www.researchgate.net/publication/331982853_Tenure_and_Faculty_Course_Evaluations).
7. Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings?. *Frontiers in Psychology*, 7, 570.
8. MacNeill, 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.
9. Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598-642.
10. Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research 2014* (DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1).

---

This material is based upon work supported by the National Science Foundation's Division of Undergraduate Education award 1654273. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.