– Best Practices for Interpretation of the Skidmore College Student Rating Form–

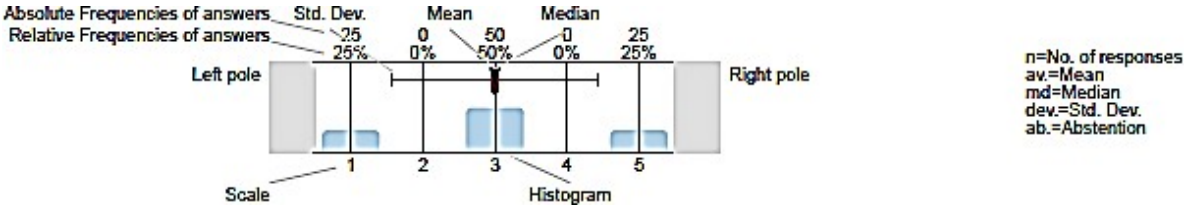
Skidmore College values highly effective teaching. However, assessing teaching it turns out is an equally challenging endeavor. Historically, Skidmore, has relied on survey of students as a primary mechanism of evaluating teaching effectiveness. Skidmore prides itself on teaching quality and on the hard work of assessing teaching effectiveness in a comprehensive manner. Acknowledging that student ratings should only a part of one's portfolio, they are included and thus, the considerations below are presented for use by faculty and faculty who are evaluating a colleague's portfolio, whether at the department, committee, or administrative level.

Considerations for Administering Forms

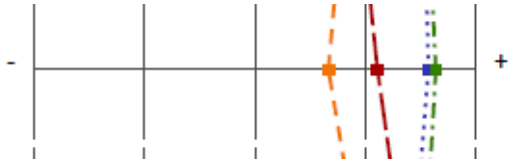
1. Students should be made aware when the forms will be administered, and/or care should be taken to ensure a majority of students are present on the pre-identified day.
2. Administration of ratings forms should not be conducted on the day of a quiz, test, or large assignment.
3. Instructors should clearly identify a student to read aloud the directions included in the packet and ensure return of the completed forms to the appropriate party. Instructors should not return the forms themselves.
4. The instructor should leave the classroom to ensure anonymity, and should give adequate time to the ratings process, the last 5 minutes before class is adjourned is not sufficient, and instructors might consider administering at the start of class. If the instructor has concerns over possible group dialogue they may ask a colleague to oversee the completion, but not return, of the ratings forms.

Considerations for Interpreting the Student Rating of Course and Instructor Summaries

1. Those reviewing the summaries of a colleague should consider the standard deviation or the variability in the measurement, which is important when considering scores. details on form and its output are available on this IR webpage. On the individual item summary for a given course item the SD is the horizontal bar pictured below.



It is critical to note that, often, small differences between a rating and a mean are not statistically meaningful. To know if a score is meaningfully different it is essential to not only compare the means, but the standard deviations as well. For example, an instructor with a rating of 4.11 on instructor overall (item 4.1) might seem lower then the college average of 4.57 for a semester. However, when considering the college SD is 0.75 the rating of 4.11 falls within the SD (mean of 4.57 – SD of 0.75 = 3.82), thus any score between 3.82 and 4.57 should not be considered meaningfully different from the College mean. This consideration is even more important in smaller courses. This should be born in mind when exploring the line plots (example below). An additional level of caution may be warranted as the data are obviously not randomly sampled, there may be concern of the SD.



n=9 av.=4.11 md=4.00 dev.=1.05
n=18 av.=3.67 md=4.00 dev.=1.24
n=308 av.=4.64 md=5.00 dev.=0.65
n=10581 av.=4.57 md=5.00 dev.=0.75

2. One should consider the number of students who completed the ratings form. The validity of data from low samples sizes can be fraught. Relatedly, one should also consider how many completed the form out of those enrolled, providing insight into the representativeness of the data. In the example below while the response rate is perfect, the number of students is still low, and more susceptible to skewness or one or two students impacting the means.

10 responses / 10 enrolled = 100% Response Rate

3. One should consider the rules of significant figures, in so far as, not reporting digits calculated beyond the precision of the original data or the precision of the instrument (e.g. 4.12345, when the form only measures in whole numbers). Consider the example of a real estate agent who sold an average of 11.621 houses/year of the last 5 years, what is 0.001 of a house? While mathematically possible, and provided in the summary, one should question the value, or significance, of these digits, and appropriate rounding should be considered.

4. One should consider that not all means are created equal. Consider the situation of two sets of ratings with a similar mean, one could be made of high with some poor scores whereas another could receive modest scores across the board. Exploring additional metrics beyond the mean, such as the median or the frequency distributions (example in item 1 above) can be useful to go beyond the known issues with means.

5. One should consider Instructor characteristics. There is relatively consistent evidence of bias against women, people of color, age, those who are non-native English speakers in higher education. In partial agreement, Skidmore found a small, but statistically significant negative effect of faculty reporting as a person of color (African, Latin, Asian, Native American). Interestingly, at Skidmore the % of the class that was female was negatively associated with student ratings. More complete student demographic data, might show the bias effect could be more significant, as shown by other studies. Finally, previous research has documented that factors unrelated to teaching, such as a greater perceived attractiveness (1, 9), gender (6), or "likability" (5) of the instructor significantly influences scores. Thus, instructors are individuals and their individual characteristics, unrelated to teaching, play into their ratings, which urges caution in any comparison to others.

6. One should consider course characteristics. At Skidmore we found that academic division, course level, course enrollment, and # of credits were found to impact student ratings. Instructors and/or department/program colleagues should provide course context that may aid interpretation.

7. There is general concern about the validity or meaningfulness of such ratings (8). Consider one study that found students who saw a silent 30 second video of the instructor and rated them, their ratings then correlated with the end of semester scores (1). Consider that other studies have demonstrated that students who perform well in a first course and rate the instructor correspondingly, perform worse in subsequent courses, or that teaching effectiveness might be

negatively correlated with student ratings (2, 3, 7) or is unrelated (8). Also, those who adopt active learning may be penalized despite greater learning (4).

8. Bearing the above in mind, the numerical ratings should be considered onto themselves, and relative to the instructor only. There should be relatively minimal concern of a singular "lower" rating in the context of an instructor's overall term or career. We all know that you can do the same thing twice and get two different results. Each class has a unique chemistry, for better or worse.

9. Last but not least, the student ratings should but a piece in an otherwise comprehensive portfolio, in terms of individual faculty presentation, but also in the reading of those portfolios when making decisions about appointments and promotions.

In summary, the above recommendations should not be considered an exhaustive list of recommendations or practices that could be evolved to better assist in our evaluation of teaching, nor be considered the end of such discussion.

References

1. **Ambady N, and Rosenthal R**. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *J Pers Soc Psychol* 64: 431-441, 1993.
2. **Braga M, Paccagnella M, and Pellizzari M**. Evaluating students' evaluations of professors. *Economics of Education Review* 41: 71-88, 2014.
3. **Carrell SE, West, James E**. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *National Bureau of Economic Research Working Paper Series* 14081: 2008.
4. **Deslauriers L, McCarty LS, Miller K, Callaghan K, and Kestin G**. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences* 116: 19251-19257, 2019.
5. **Feistauer D, and Richter T**. Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation* 59: 168-178, 2018.
6. **MacNell L, Driscoll A, and Hunt AN**. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education* 40: 291-303, 2015.
7. **Stroebe W**. Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations. *Perspectives on psychological science : a journal of the Association for Psychological Science* 11: 800-816, 2016.
8. **Uttl B, White CA, and Gonzalez DW**. 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, 2017.
9. **Wolbring T, and Riordan P**. How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching. *Soc Sci Res* 57: 253-272, 2016.