An Interpretation Guide for the

Student Opinion of Teaching Effectiveness Surveys (SOTES)

Prepared by:

The SJSU Student Evaluation Review Board (SERB)

The information presented here includes a description of the SOTE instrument, and overview of the statistics included in the SOTE report, and a brief review of factors that influence SOTE ratings. Note that the language of an interpretation guide is not policy but primarily factual information (F04-1).

[Rev Fall 2024]

Table of Contents

SOTE Interpretation Quick Guide

Background and Administration

- The Student Evaluation of Teaching Effectiveness (SOTE) instrument was created to assess student perceptions of teaching effectiveness (the current version was revised in Fall 2019).
- The survey begins with a brief introduction and overview, followed by 13 closed-ended items, 4 informational items, and 3 open-ended questions.
- SOTE surveys are administered by the SJSU Office of Institutional Research and Strategic Analytics (IRSA) through CourseEval (online software integrated with Canvas).

Results, Reporting, and Interpretation

- Results are provided to individual instructors and department chairs. Results are also included in annual and cumulative evaluations for lecturers as well as faculty evaluations for retention, tenure, and promotion (RTP).
- Reports include means, medians, standard deviations, and percentile rankings for the instructor as well as norms for instructor's department, college, and the university as a whole.
- Ratings should be considered atypical or extraordinary only when they fall outside the reported norms (20-80th percentile range). Interpretation should take into account class size, response rate, and trends across classes and semesters.
- Evaluations of student responses to open-ended questions should consider the totality of comments (rather than focusing on individual comments).
- While responses to Question 13 are often used as an index of overall effectiveness, evaluations of teaching effectiveness should be based on results for all questions.
- Alongside the Collective Bargaining Agreement, University Policy F12-6, "Evaluation of Effectiveness of Teaching for All Faculty" establishes that SOTEs/SOLATES may not be the only form of evaluating academic assignment and they should be assessed in a broader context. Thus, SERB recommends that RTP committees use SOTE ratings as just one metric by which to evaluate instructor effectiveness.
- Several factors are known to systematically influence student evaluations, including academic discipline, course level, class size, student grades, and instructor characteristics (e.g., gender, race and ethnicity, and language background).

Relevant Policy

- Instructors may request the removal of student remarks that are completely unrelated to teaching (e.g., comments that are bigoted, hateful, evaluate personal appearance, or otherwise violate campus policies).
- Faculty may occasionally exclude the results of up to one course per academic year from their periodic evaluations (provided they teach at least fifteen units during that Academic Year).
- In response to the COVID-19 pandemic, University Policy S20-4 allows for faculty members to exclude SOTEs administered during Spring 2020 if they so choose. No negative inferences should be drawn if faculty elect to exclude Spring 2020 SOTEs.

- A recent memo from Provost Del Casino also required that RTP committee members "interpret SOTEs from Fall 2020 with care" due to the challenges of converting instruction modality.
- Instructors and department chairs may request a report of responses to questions asking about 'undue influence' from the IR Office. Typically, such requests occur when students make independent allegations of improprieties and an investigation is conducted.

Questions? For an up-to-date listing of Student Evaluation Review Board members (which includes one representative per college), visit

https://www.sjsu.edu/senate/committee-taskforce-information/assignments.php

History and Policy

The Student Evaluation Review Board is an Operating Committee of the Academic Senate that reports to the Professional Standards Committee. The board includes one faculty member from each of the seven colleges on campus as well as one student representative (at-large). The directors of the Office of Institutional Research and Strategic Analytics (IRSA) and the Center for Faculty Development serve as ex officio members on the committee.

The committee is charged with designing evaluation instruments to be used by all departments and colleges, developing guidelines for the participation of students in the evaluation of faculty, and reviewing proposals for matters concerned with rating instruments, norm grouping or any other variance to established policy.

In addition, SERB is charged with constructing and establishing norms for the rating instruments such that an instructor's ratings can be compared with average ratings of colleagues teaching similar courses across the university. This Interpretation Guide was created to provide information and guidelines for the effective interpretation of the rating instruments, thereby making it possible to form a better judgment about an instructor's teaching effectiveness.

Provost Vincent Del Casino issued a memo **"Guidance for RTP and Lecturer Evaluations in the Era of Pandemic"** on August 10, 2020. In Section B, "Teaching-Related Considerations", Provost Del Casino made several recommendations with respect to SOTEs.

- "1) Draw no negative inferences if faculty elect to exclude Spring 2020 SOTEs. Faculty are allowed (by S20-4) to exclude the results of SOTEs conducted during Spring 2020 from their "Working Personnel Action Files" (materials submitted for performance reviews such as dossiers). Faculty are also routinely allowed to exclude the results of other SOTEs (approximately 1 per year for faculty who meet this exception under F12-6, E4) from their evaluation process."
- "2) Interpret SOTEs from Fall 2020 with care. Many faculty were teaching semester-long online courses for the first time. Some courses are extremely difficult to convert to an online modality and some students dislike online modalities. Students could also voice negative reactions that have little to do with the quality of the instructor's efforts or the instructor's ability. Evaluators must read the entire SOTE and contextualize the differences that faculty may see in these relative to other similar courses taught in different modalities. Reviewers should carefully review all the SOTE measures, both quantitative and qualitative."
- "3) Contextualize teaching with a holistic view. SJSU policy says: "When evaluating effectiveness in teaching, chairs, committees, and administrators are required to conduct a holistic evaluation. This means that teaching must be considered in context and must be evaluated using multiple sources of information" (F12-6). The COVID-19 pandemic is a paramount contextual factor when evaluating teaching conducted beginning Spring 2020."
- "4) Policy prohibits reliance solely on SOTEs to evaluate teaching. During the current climate, it is even more important to evaluate teaching success in the context of the unfavorable conditions created by the pandemic."

The following overview highlights some key policies related to SOTE administration and interpretation. For a complete index of SOTE policies, visit https://www.sjsu.edu/senate/university-policies/policies-by-category/policy-sote.php.

F12-6: When evaluating effectiveness in teaching, chairs, committees, and administrators are required to conduct a holistic evaluation. This means that teaching must be considered in context and must be evaluated

using multiple sources of information [including context, purpose, and course objectives, implementation of the course, and direct observation by peers].
F12-6: Since student opinion surveys measure student satisfaction rather than student learning, they cannot be considered perfect indicators of teaching quality…. To guard against the limitations of the instrument, all those using SOTES as part of the SJSU evaluation process must consult the official interpretation guide… Information from SOTES is but one source of information for assessing teaching effectiveness.
F12-6: SOTES shall be administered in all classes [except those officially excluded for technical or ethical reasons] and the results placed in the faculty personnel file. Faculty, however, under some circumstances may exclude the results of an occasional course from their periodic evaluations. Faculty may choose to exclude the survey results from one course per Academic Year from their periodic evaluation, provided that they teach at least fifteen units of courses during that Academic Year.
F12-6: Any SOTE with a response rate of less than 50% or with fewer than 10 responses will be flagged as potentially unreliable and interpreted with caution.
F12-6: Faculty may request the removal of remarks in the qualitative surveys that are completely unrelated to teaching, such as comments that are bigoted, hateful, comment on personal appearance, or otherwise violate campus policies. Such remarks will be removed after verification of their content by the Department Chair.
F12-6: Results shall be reported as the means, standard deviations, and medians for each item by class. The mean for each class will be compared against the mean and norms for the particular College and University when appropriate. The frequencies of responses (e.g., the number of "5"s and "4"s and "3"s etc.) for each question will also be reported.
F12-6: Norms (an indicator of the middle range of scores) shall be provided to assist in the interpretation of quantitative SOTES.
F12-6: SOTES shall be collected by electronic means. The AVP for IEA shall arrange for all students to receive regular electronic reminders to complete their SOTES, and these reminders will inform students how to connect to and complete the survey instrument…. Statements that clearly explain to students the seriousness with which SJSU takes the results of the survey… should be provided both in the electronic reminders and at the beginning of the survey instrument.
F12-6: SOTES shall not be [administered] earlier than the final 10 days for class nor later than the normal time when the student's final grade is released. A minimum of 10 calendar days will be provided to respond. No SOTE results… may be released to faculty until after grades for the class are officially submitted. No students will be allowed to submit SOTES after they have seen their official grade for a course.
S14-1: Amendment to F12-6 "Evaluation of Effectiveness in Teaching for All Faculty." Under some circumstances faculty may exclude the results of an occasional course from their periodic evaluations. Faculty may choose to exclude the survey results from one course per Academic Year from their periodic evaluation, provided that they teach at least fifteen units of courses during that Academic Year. Faculty who are credited with teaching double sized courses will be credited with teaching twice the normal number of units.
S17-2: The revised versions of the SOTE and SOLATE questionnaires were approved and deemed effective for the administration as soon as possible.
S15-8 Amendment B: Retention, Tenure and Promotion for Regular Faculty Employees: Criteria and Standards revises "Baseline" criteria for academic assignment. "3.3.1.3.2 Baseline. The candidate has documented effectiveness in teaching, particularly for classes within the candidate's primary focus and any curriculum specifically identified in the appointment letter. Assigned courses are well crafted and appropriate for the catalog description, as shown in course syllabi and other teaching materials. The candidate has taken measures

to correct any problems identified earlier in either direct observations or prior performance evaluations. Recent direct observations and surveys of student opinion of teaching effectiveness (SOTEs) are also supportive. SOTEs are considered supportive if they are either within appropriate norms, or if a preponderance of student opinion from objective and subjective questions indicates effective teaching."

S20-4: Optional Exclusion of Student Opinion of Teaching Effectiveness surveys (SOTEs) Administered during Spring 2020. Faculty be permitted, at their option, to exclude any SOTE results obtained during Spring 2020 from future evaluations.

S20-7: Students were allowed to petition the Registrar to change a letter grade to Credit/No Credit for all classes. Spring 2020 SOTE results incorrectly excluded students who petitioned the Registrar to change a letter grade to Credit/No Credit.

The SOTE Survey

The most recent version of the SOTE instrument was administered for the first time in Fall 2017. See below for a comparison across the old and new instruments. Note that both versions begin with a brief introduction and overview, followed by thirteen (13) closed-ended items that assess students' perceptions on teaching effectiveness and their learning experiences. These are followed by four (4) informational items and three (3) open-ended questions. Items and instructions that were revised in Fall 2017 are in **bold font**.

Instructions

This instrument is designed **to be a professional evaluation** of your instructor's teaching performance. It is NOT designed to measure your reaction to the subject, the facilities (such as the physical conditions of the classroom), **or your instructor's physical appearance**. Your individual ratings will be anonymous and a summary of items 1-18 will be available to your instructor after grades are turned in. This summary may enhance your instructor's teaching. It will also be used in the evaluation of your instructor for personnel matters such as retention, tenure and promotion**. If the question does not apply to your course, please select "not applicable/no opportunity to observe"**.

Closed-Ended Questions

Topic	**Item**	**Old** *(Fall 2003 – Spring 2017)*	**New** *(Fall 2017 – present)*
Relevance	Q1	Demonstrated relevance of the course content.	*[no change]*
Learning Environment	Q2	Used assignments that enhanced learning.	*[no change]*
Helping Students Think	Q3	Summarized/emphasized important points.	*[no change]*
Learning Environment	Q4	Was responsive to questions and comments from students.	*[no change]*
Learning Environment	Q5	Established an atmosphere that facilitated learning.	*[no change]*
Responsiveness to Students	Q6	Was approachable for assistance.	*[no change]*
Responsiveness to Students	Q7	Was responsive to the diversity of students in class.	Was **respectful of** the diversity of students in class.
Learning Environment	Q8	Showed strong interest in teaching this class.	*[no change]*
Helping Students Think	Q9	Used intellectually challenging teaching methods,	**Used teaching methods that helped students learn important concepts.**
Grading and Feedback	Q10	Used fair grading methods.	**Used grading criteria that were clear.**
Helping Students Think	Q11	Helped students analyze complex/abstract ideas.	*[no change]*
Grading and Feedback	Q12	Provided meaningful feedback about student work.	*[no change]*

Overall Effectiveness	Q13	Overall, this instructors teaching was: *(5, very effective; 4, effective; 3, somewhat effective; 2, ineffective; 1, very ineffective)*	Overall, this instructor's teaching was **effective**.

Notes: Items and instructions that were revised in Fall 2017 are in **bold font**. Response options for Questions 1-12 on the old instrument (Fall 2003 - Spring 2017) used the following scale: *5, Very Strongly Agree; 4, Strongly Agree; 3, Agree; 2, Disagree; 1, Strongly Disagree; NA, Not Applicable/No Opportunity to Observe*. The new instrument (Fall 2017 - *present*) adopts a slightly modified scale (for all questions): *5, Strongly Agree; 4, Agree; 3, Neutral; 2, Disagree; 1, Strongly Disagree; NA, Not Applicable/No Opportunity to Observe.*

Informational Questions

Item	**Old** *(Fall 2003 – Spring 2017)*	**New** *(Fall 2017 – present)*
Q14	What is your current estimate of your expected overall grade in this course? *(A; B; C; D or F; Other)*	*[no change]*
Q15	You are a: *(Freshman; Sophomore; Junior; Senior; Graduate Student; Credential Student; Other)*	*[no change]*
Q16	Did you complete this form without undue influence from other students? *(Yes; No)*	*[no change]*
Q17	Did you complete this form without undue influence from the instructor? *(Yes; No)*	*[no change]*

Open-Ended Questions

Item	**Old** *(Fall 2003 – Spring 2017)*	**New** *(Fall 2017 – present)*
Q18	Discuss the strengths of this instructor's teaching.	**What do you think are the strengths of this instructor's teaching?**
Q19	Discuss the weaknesses and/or areas in need of improvement of this instructor's teaching.	**What suggestions, if any, do you have to further improve the instructor's teaching?**

Q20	Please provide any other comments you feel would be helpful to the instructor regarding his/her teaching performance/ability.	**If you like, please use this space to elaborate on your responses.**

Interpretation of the SOTE Ratings

SOTE Reporting

To aid in interpretation, official SOTE reports provide data (means, standard deviations, and medians) for the instructor's department, college, and the university as a whole.

- **Mean:** This is the arithmetic average of student responses. Note, however, that most student rating distributions are skewed (that is, the ratings bunch up toward one end, typically the right end), in which case the mean does not represent the typical or most frequently occurring rating. Experts recommend avoiding the use of the mean for reporting central tendency of ordinal data; the median is a more appropriate indicator for identifying the central tendency of SOTE data.
- **Standard Deviation:** This statistic measures the variability among the responses (i.e., how much, on the average, student responses vary from the mean). Like the mean, the standard deviation is an inappropriate measure of variability when the response data is ordinal.
- **Median:** This is the middle ranking. A median of 3.5 indicates that half the students gave ratings higher and half lower than 3.5. The median is helpful in cases where outliers might influence the mean and standard deviation (e.g. cases in which a few extremely high or extremely low ratings push the mean score in a direction that is not representative of the class as a whole). This is particularly likely in smaller classes or classes with large numbers of blank or "not applicable" ratings.
- **Norms:** Norms reported via the CoursEval system are updated each semester. In addition to the statistics mentioned above, reports to faculty include the exact percentile of the faculty's mean score relative to department and university norms (college norms are also reported as supplemental material)[1]. These percentiles can be used to compare an instructor's ratings with the average ratings of colleagues. Consistent with previous interpretation guidelines, ***percentile rankings within the 20-80 range should not be interpreted as anything other than typical since they are close to the median. Ratings that fall outside this range (below 20 or above 80) should be interpreted as the top 20% and bottom 20% of the sample. It is sometimes customary to consider the top or bottom 10% and 5% extreme, but those values are arbitrary. The definition of atypical needs to be based on more rigorous analysis using data triangulation with other sources. Further, the interpretation of these results should be done using trends across classes and semesters.*** If the mean response to any particular question is ***consistently*** below (or above) the norm then the item should be noted as important. RTP Committees should make special note of S15-8 Amendment B which revises the "Baseline" criteria for academic assignment. "… [r]ecent direct observations and surveys of student opinion of teaching effectiveness (SOTEs) are also supportive. SOTEs are considered supportive if they are either within appropriate norms, or if a preponderance of student opinion from objective and subjective questions indicates effective teaching."[2]
- **Open-Ended Responses:** Students' written comments provide additional information on teaching effectiveness. In interpreting these responses, members of RTP committees should take into account the majority of comments, rather than focusing on individual responses. However, if comments are

[1] The old reporting format (Fall 2003 – Spring 2017) indicated the middle 60% of ratings received by instructors for each college, and for the university as a whole, as a line of dashes. The instructor's mean for this course was indicated by an asterisk on the same line.

[2] SERB recognizes the use of norms is problematic and should be critically reassessed. Experts such as Stark have sharply criticized norms. SERB will consider an " addendum" to the next version of this guide in 2026.

repeatedly observed for the same instructor, then RTP committees should consider further evaluations for that instructor.

Overview of Reliability[3]

The norms and statistics reported in this Interpretation Guide were calculated from SOTE survey results from Fall 2022 and Spring 2023. All courses across all colleges were included in this analysis, resulting in a total of 117,864 student responses (Fall '22 = 62,248 responses; Spring '23 = 55,616 responses). IRSA has provided the calculated the reliability coefficients for this data.

Typically, the reliability of an instrument refers to the degree to which an instrument's scores for a group of respondents are consistent over repeated applications of a measurement procedure." (AERA, APA, NCME 2014, p. 223). An instrument's reliability can also be defined as the degree to which it is internally consistent (i.e., the degree to which items correlate to identify a dimension or construct). A recent study in Fall 2022-Spring 2023 found that Cronbach's alpha (α) was 0.97 across all 13 questions, indicating a very strong level of internal consistency across SOTE fixed choice questions.[4] An alpha (α) of .92 has been reported in a previous study of the SOTE instrument's reliability (See SERB Technical Report).[5] For survey data, relatively high indicators of internal consistency are common but not conclusive (Sijtsma, 2015; Maul, 2017).

We also note that Question 13 is strongly correlated with all of the other items. While Question 13 is often used as an index of overall effectiveness, we recommend that evaluations of teaching effectiveness consider all 13 items.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
Q1	1	0.738	0.758	0.655	0.720	0.639	0.622	0.697	0.719	0.623	0.723	0.637	0.749
Q2		1	0.760	0.641	0.733	0.636	0.567	0.644	0.782	0.651	0.745	0.684	0.786
Q3			1	0.694	0.761	0.670	0.607	0.695	0.790	0.660	0.785	0.687	0.802
Q4				1	0.761	0.809	0.654	0.684	0.683	0.634	0.708	0.691	0.739
Q5					1	0.763	0.666	0.741	0.801	0.665	0.791	0.718	0.829
Q6						1	0.690	0.693	0.686	0.636	0.715	0.696	0.741
Q7							1	0.671	0.592	0.572	0.627	0.578	0.631
Q8								1	0.710	0.601	0.719	0.656	0.737
Q9									1	0.687	0.832	0.736	0.861
Q10										1	0.710	0.694	0.726
Q11											1	0.764	0.850

[3] There can be no meaningful reliability without validity. The instrument is quite vulnerable if there is no substantial validity evidence to support its consequential uses for RTP. This issue will be taken up by SERB in 2025.

[4] Maul (2017) notes: The results…suggest that, at least in the context of responding to survey questions, respondents often choose to behave consistently unless there is a clear reason not to do so which could be interpreted as an extreme, limiting case of 'self-generated validity,' in the terminology of Feldman & Lynch, 1988). As such, it may be that favorable-looking results of covariance-based 10 statistical procedures (such as high reliability estimates and fit to unidimensional latent variable models) should be regarded more as a default expectation for survey response data than as positive evidence for the validity of an instrument as a measure of a psychological attribute." (p. 8) [Italics added]

[5] Duckor, B., Chen, C., Currin-Percival, M., Tortora, C., & Villagran, M. (2024, February). *SOTE instrument technical report*. Student Evaluation Review Board Committee, San José State University. [Technical Report]. https://drive.google.com/file/d/1WWhIZ2lZ-sKiCbTMv89D_e4UNfifJPMM/view?usp=sharing

Q12												1	0.790
Q13													1

The Pearson product moment correlation measures the strength of linear dependence between two variables, and varies between -1 and 1. As a rule of thumb, correlations between .00 and .50 are considered weak; correlations between .50 and .70 are moderate, and correlations over .70 are relatively strong. The correlations presented in the table above are all statistically significant at the p<.01 level.

In Fall 2022, 5.2% of students (n=3,229) responded 'no' to Question 16 ("Did you complete this form without undue influence from other **students**?") and 5.4% of students (n=3,316) responded 'no' to Question 17 ("Did you complete this form without undue influence from the **instructor**?"). Of these students, most (n=3,017) responded 'no' to both questions indicating that they may have misunderstood the question. In Spring 2023, 5.2% of students (n=2,860) responded 'no' to Question 16 ("Did you complete this form without undue influence from other **students**?") and 5.3% of students (n=2,945) responded 'no' to Question 17 ("Did you complete this form without undue influence from the **instructor**?"). Of these students, most (n=2,691) responded 'no' to both questions indicating that they may have misunderstood the question.

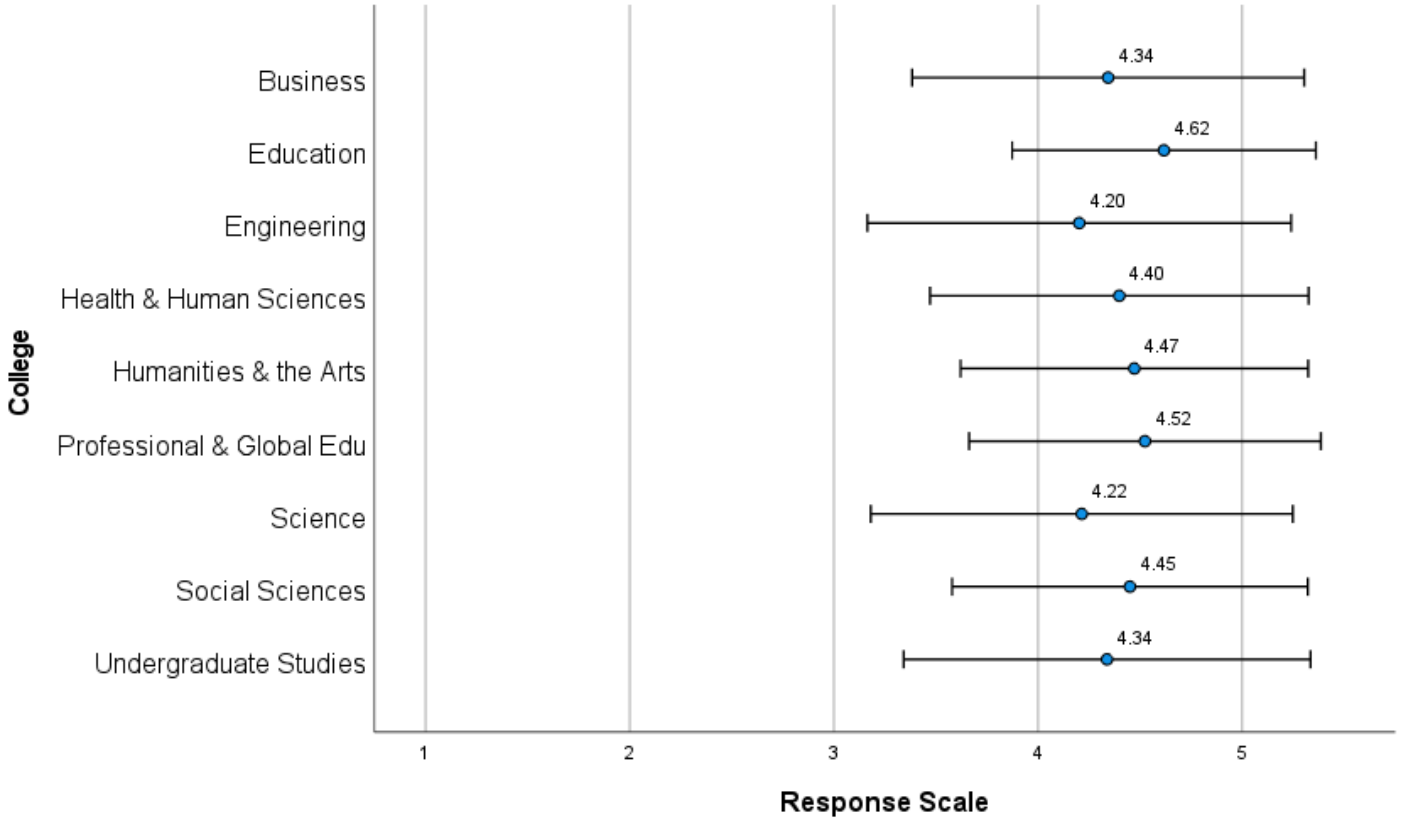
We also note that ***several factors are known to systematically influence SOTE ratings***. This is demonstrated below using Fall 2022-Spring 2023 data with references to similar findings from research conducted elsewhere. These factors should be considered in any RTP evaluation of SOTE data and we encourage faculty members to include additional information and explanation in their dossiers as necessary.

Course Characteristics

College and Content

From the following figure, there appear to be some differences in the averages of the ratings of overall teaching effectiveness (Q13) across colleges at San Jose State. This is a common trend, Stroebe 2020 contains a review of articles studying the correlation between disciplines and faculty ratings, with faculty in science and engineering obtaining less positive ratings (see also Gravestock & Gregor-Greenleaf, 2008).

Mean Response of Instructor's Teaching Effectiveness by College



Error Bars = +/- 1 SD

There are also differences in average ratings between departments within colleges. It is therefore important that RTP committees evaluating candidates from different departments and colleges (College and University level RTP committees) compare instructors to colleagues within their own departments and colleges in addition to the overall university.

Research on student evaluations at other universities has also shown that ratings are often lower when students are required to take a class as compared to when they are taking the class as an elective (Arreola, 2000). Similarly, students often offer higher ratings to courses outside their area of study than to courses within their major (Theall & Franklin, 2001). Moreover, class size and course level may have an impact (Gravestock &

Gregor-Greenleaf, 2008). Kreitzer et al. 2022 identify all the mentioned characteristics as measurement bias, when variables unrelated to teaching effectiveness systematically influence the results, and provide a wide literature review. Note, however, Beran et al. (2009) argue that these effects may be mediated by varying levels of student engagement.

Innovative Pedagogy

Significant differences in student evaluations are observed due to course type and pedagogical structure, which can be daunting for faculty engaging in pedagogical innovation to improve student learning and lead to an entrenchment of traditional lecture-heavy, teacher-centered pedagogies. Numerous studies have cautioned against using student evaluations as an indicator of student learning, with student learning outcomes explaining only 1-14% of the variability in student evaluations (e.g., Uttl, White & Gonzalez, 2017; Clayson, 2009; Cohen, 1980). In addition, student evaluations were generally developed to assess a teacher-centered learning environment, with a knowledge transmission model, and in many cases fail to capture the benefits of desettling the classroom to a student-centered paradigm (Kolitch & Dean, 1999; Theall, 2010). Kember, Jenkins & Ng (2010) argue that student responses on evaluations depend largely on what students consider to be good teaching; this may align more with student previous experience or disciplinary conceptualizations rather than effective teaching practice.

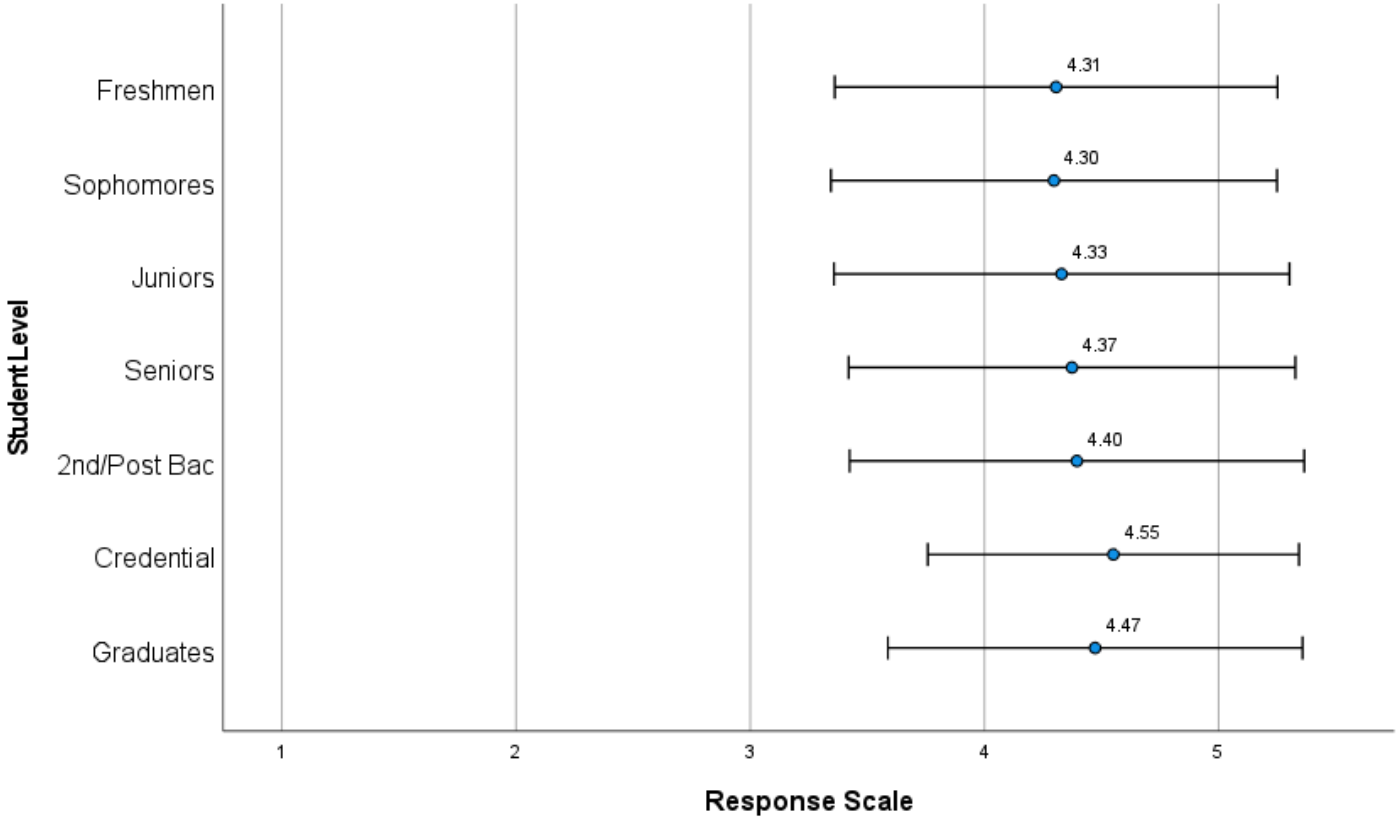
While some studies have shown a positive correlation between innovative approaches like the flipped classroom model and student evaluations (Samuel, 2019; Lag & Saele, 2019), this is not universally the case with pedagogical innovation and teaching effectiveness. The fear of decreased evaluations and the consequences therein can lead faculty to be wary of negative consequences of pedagogical innovations that are research-based as these can lead to lower student evaluations (Henderson, Khan & Dancy, 2018). It should be emphasized that student evaluations are assessments of student experience, not of student learning, and many researchers have questioned whether students have the ability to assess the appropriateness and effectiveness of the pedagogies faculty employ as student evaluations and other metrics of teaching effectiveness are often anti-correlated (Braga, Paccagnella & Pellizzari, 2014; Kornell & Hausman, 2016). This leads both to the correlation between perceived easiness (and thus grades obtained, as discussed below) and student evaluations, as well as a decrease in student evaluations when students experience discomfort (Felton, Mitchell & Stinson, 2004; Walker et al., 2008).

In many innovative, student-centered approaches, the burden of knowledge construction is more clearly placed on the student, which can lead to student unease as they may be concerned about not having a clear-cut correct answer, of being negatively evaluated, or experiencing greater anxiety due an increasingly active role and responsibility in the class (Cooper, Downing & Brownell, 2018; Downing et al., 2020). This discomfort can lead students to evaluate an instructor less favorably, even when student performance increases (Walker et al., 2008), or if faculty are employing active learning strategies that have been demonstrated to be more effective (Hake, 1998; Freeman et al., 2014). While there are strategies to ameliorate this discomfort (Cooper, Downing & Brownell, 2018), it is unlikely that a faculty member trying out new pedagogical approaches and innovating in their teaching would be fluent in navigating these challenges. In fact, student evaluations may penalize the most innovative faculty who are supporting greater student learning and performance in their courses (Walker et al, 2008; Braga, Paccagnella & Pellizzari, 2014; Kornell & Hausman, 2016).

Course Level

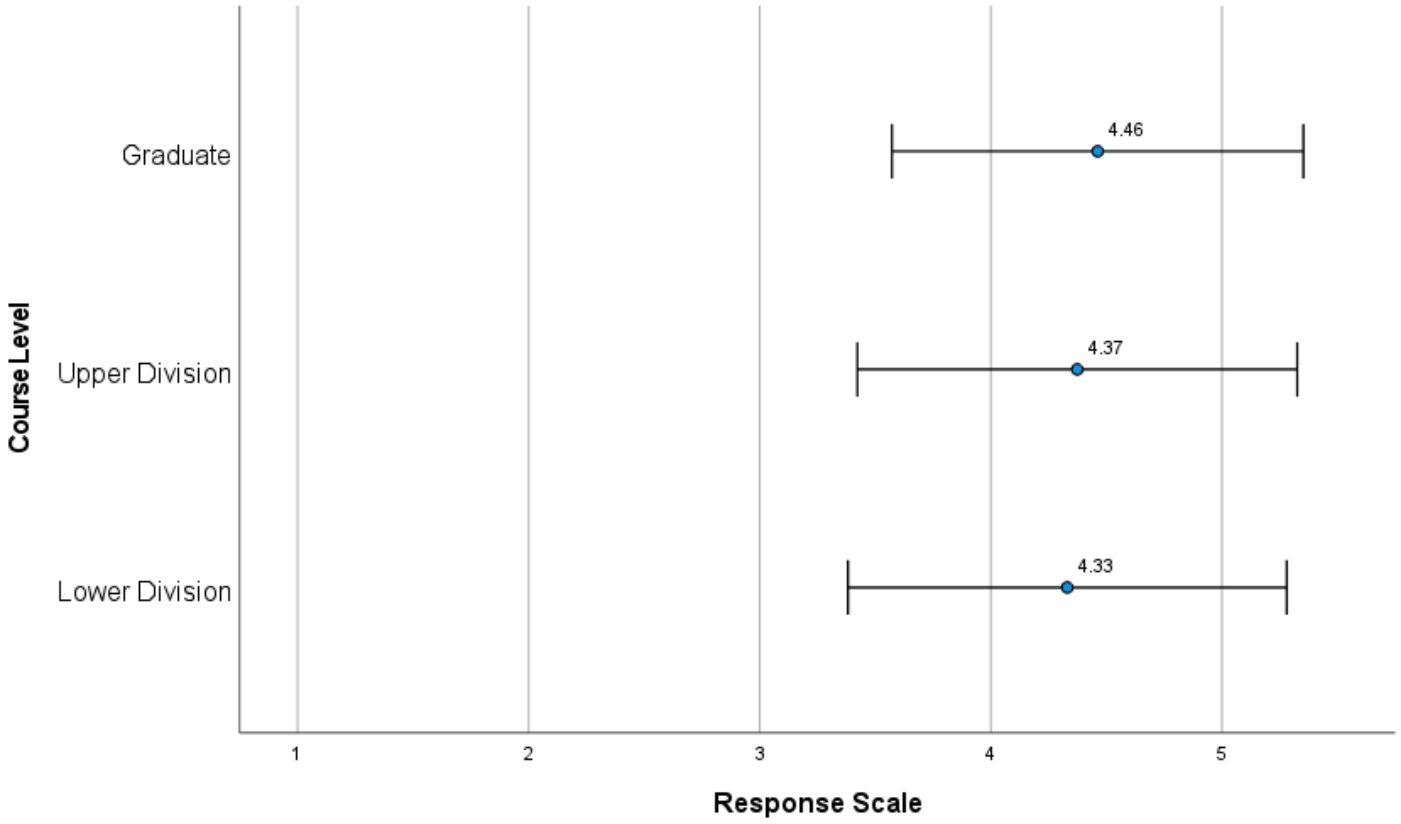
There appear to be slight differences in the average ratings of overall teaching effectiveness (Q13) across student level (i.e., frosh, junior, graduate, etc.) as well as level of instruction (e.g., upper- vs. lower-division courses).

Mean Response of Instructor's Teaching Effectiveness by Student Level



Error Bars = +/- 1 SD

Mean Response of Instructor's Teaching Effectiveness by Course Level



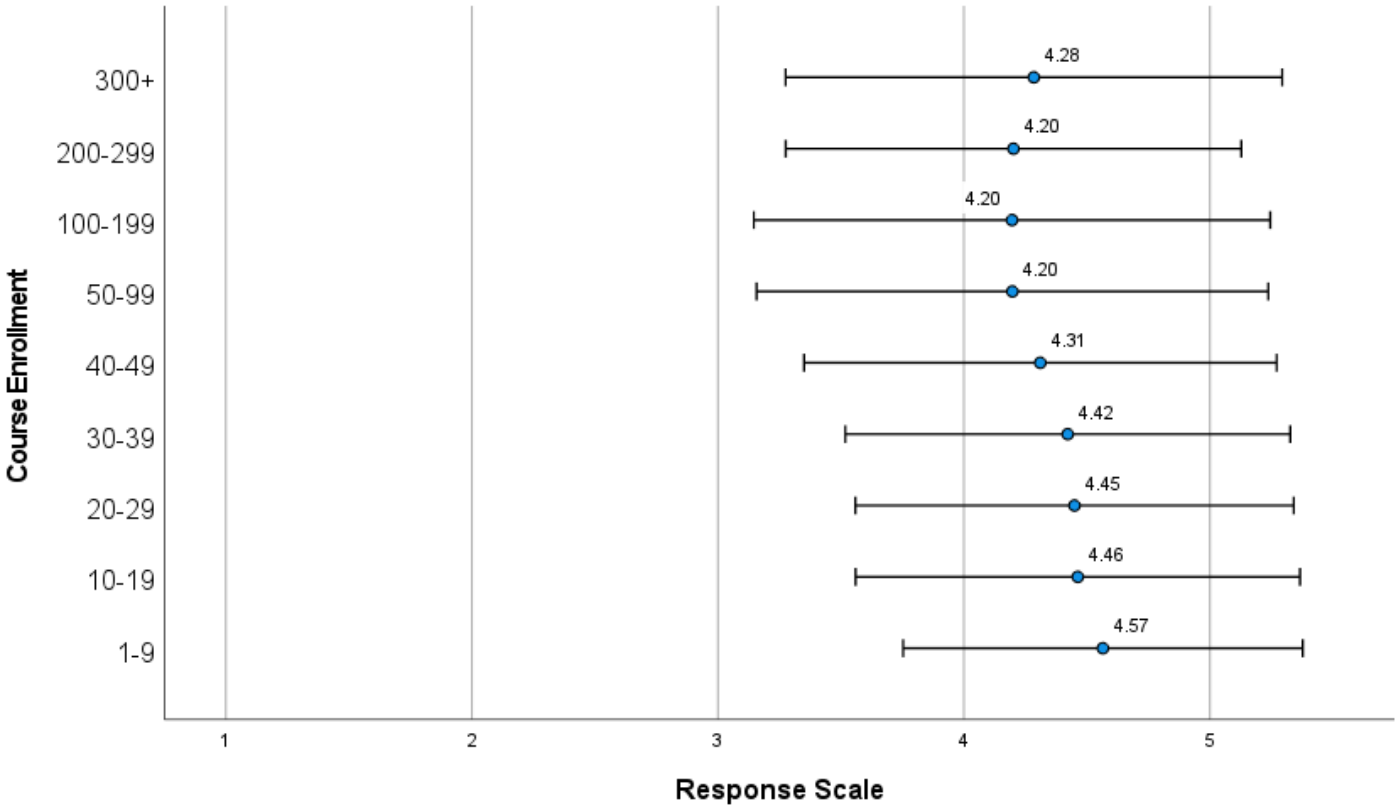
Error Bars = +/- 1 SD

Research on student evaluations at other universities shows that ratings in graduate and credential classes tend to be higher than in undergraduate classes (see also Arreola, 2000; Marsh & Hocevar, 1991). However, ratings across lower and upper division courses tend to be relatively similar (Arreola, 2000).

Class Size

Class size also seems to influence average ratings of overall teaching effectiveness (Q13) (Mandel & Sussmuth, 2011; Marsh, Overall, & Kesler, 1979; Marsh & Roche, 1997;). Note that class size should not be confused with the number of survey respondents or average daily attendance. Here, we consider class enrollment.

Mean Response of Instructor's Teaching Effectiveness by Course Enrollment



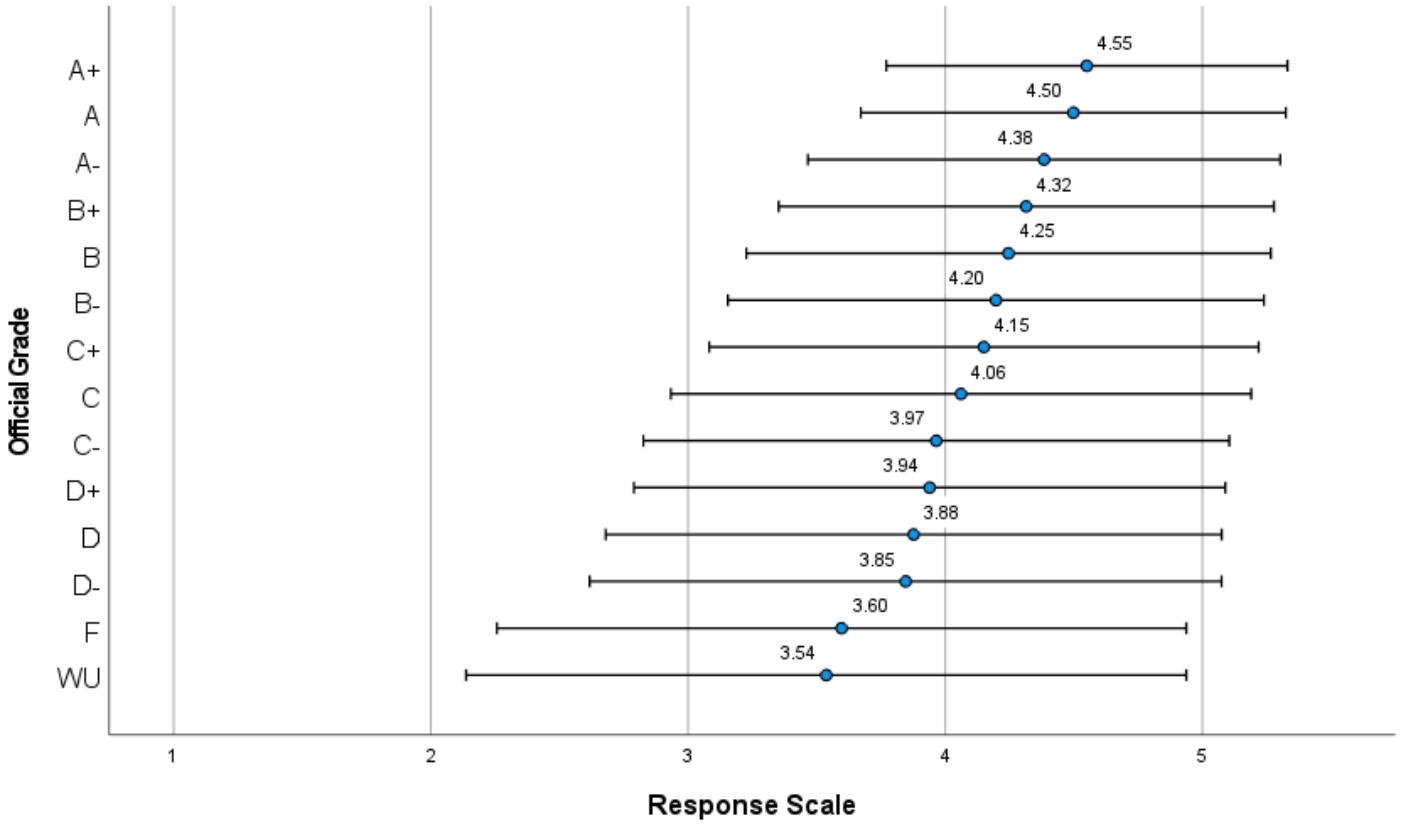
Error Bars = +/- 1 SD

Previous research has also reported a relationship between class size and student evaluations, with small or moderate sized classes (<30) rated more favorably than larger classes (Johnson et al., 2013; Mateo & Fernandez, 1996). Furthermore, Chapman and Ludlow (2010) found that increased class size (beyond 30 students) has a negative effect on "perceived student learning," a composite measure based on student self-evaluations of their own learning. Cuseo (2007) found large class size reduces the frequency and quality of instructor interaction with and feedback to students. Further students gave lower overall evaluations for course instruction delivered in larger classes. Large class size is a variable that adversely affects student learning by lowering students' active involvement with the instructor and the subject matter (p. 10).

Official and Expected Grades

Possibly the most notable impact on student ratings is their anticipated and official grade in the course.

Mean Response of Instructor's Teaching Effectiveness by Students' Official Grade[6]



Error Bars = +/- 1 SD

In fact, it is well established that student ratings are positively associated with both expected and actual course grades (e.g., Kulik, 2001). Greenwald & Gillmore (1997) further concluded that grading leniency exerts an important influence on ratings. However, another possible explanation for this result is that strong instructors teach courses in which students both learn a lot (therefore, they earn and deserve high grades) and give appropriately high ratings to the course and the instructor (Spooren and Mortelmans, 2006).

Nevertheless, when interpreting SOTE ratings, we encourage RTP committees to note the distribution of expected grades. Classes in which the majority of students expect either low or high grades should be fairly rare (exceptions to this would be graduate and credential classes in which a grade lower than a "B" is often considered equivalent to a failing grade). In addition, expected grades for a class should show some relationship to actual grades. In cases where there is a wide discrepancy (e.g. 80% of the class expects a grade of "A" while the actual average grade for the class is a "C") RTP committees may request further information from the instructor.

[6] Please note that although WUs appear on this graph, students with a WU are removed from the SOTE/SOLATE reports.

Administration

Several studies have failed to detect a significant difference in ratings between online evaluations and paper evaluations (Donovan et al., 2006; Hardy, 2003; Heath et al., 2007; Laubsch, 2006; Spooner et al., 1999). At SJSU, a study by Sujitparapitaya and Briggs (2010) indicated that there was no significant difference for a majority of the responses between online evaluations and paper evaluations (similar to findings from a study conducted at Brigham Young University, Sorenson & Johnson, 2006). While some studies have found that specific questions may be answered more favorably in online evaluations (Liu, 2006; see also Avery et al., 2006; Cao et al., 2007), others have reported that paper evaluations produced higher scores for individual questions and total scores (Chang, 2003; Mau et al., 2012).

Importantly, the overall response rate at SJSU has remained the same, if not improved, since the university moved to online implementation in 2013 (48.1% in Fall 2022; 47.1% in Spring 2023). We also note that there is no evidence for a significant difference in student responses to Question 13 across the Fall and Spring semesters ($M_{fall} = 4.36$, $SD_{fall} = .95$; $M_{spring} = 4.39$, $SD_{spring} = .93$).

A study by Guder and Maliaris (2013) showed that the response rate of online evaluation increased when emails were sent and when faculty emphasized the importance of completing the evaluations in class. Van Mol (2017) suggested that sending extra reminders with specific reminder content is effective for increasing online evaluation response rates.

Instructor Characteristics

Whereas analyses of SOTES responses in relation to various instructor characteristics are not reported here, the factors discussed below have been identified in existing literature as possible threats to the validity of student evaluations. Note that this is not intended to be a comprehensive review of such factors, but a brief review is presented here as a point of consideration.

Gender

In recent research, Mitchell and Martin (2018) analyzed student evaluations of two identical online courses – one was assigned a female instructor and the other a male instructor. They found that the male instructor was rated more favorably than the female instructor on all items in the student evaluations, even those that the instructor has no control over, such as the university registration procedure (see also Arbuckle & Williams, 2003; Chávez & Mitchell 2020; MacNell et al., 2014).

Gender role beliefs are another important factor. Students expect male instructors to be more authoritative and expect female instructors to be more nurturing, with stronger interpersonal skills (Anderson & Miller 1997; see also Mitchell & Martin, 2018). Students reward instructors who follow these gender roles (Andersen & Miller, 1997) and are more critical of those that do not (Basow et al., 2006; Chamberlin & Hickey, 2001; Dalmia et al.,2005; MacNell et al., 2014; Sprague & Massoni, 2005). For instance, Basow and Montgomery (2005) found that female professors received higher ratings than male professors on interpersonal questions and on items about faculty-student interactions (see also Bachen et al., 1999; Basow & Montgomery, 2005; Centra & Gaubatz, 2000).

Many have also found significant differences in evaluations of female and male instructors depending on the gender of the student. For example, male students often rate male instructors higher than female instructors, whereas female students rate female instructors higher than male instructors (Basow 1995; Centra & Gaubatz, 2000; Mengel, Sauermann, & Zölitz 2019). Kohn and Hatfield (2006), however, found that female students rated male faculty even higher than their male classmates.

Additional research shows other differences potentially connected to gender bias. Sinclair and Kunda (2000), for example, found that low grades negatively affect ratings that students give to female instructors, but not male instructors. Martin (2016) found an interaction between faculty gender and class size with female faculty members receiving lower evaluations in larger courses than male faculty.

Race and Ethnicity[7]

Research on the effect of race and ethnicity on student evaluations is limited. Nevertheless, there is some clear evidence that African American and Hispanic faculty members receive lower evaluations than white and Asian faculty members (e.g., Basow, Codos, & Martin, 2013). Similar lines of research have found that African American faculty members are rated lower than Caucasian faculty members on broad evaluations of teaching effectiveness (Smith, 2007; Smith & Hawkins, 2011; Smith & Johnson-Bailey, 2011).

[7] SERB is aiming to develop a Qualitative Technical report which includes a qualitative analysis of SOTES from SJSU to offer corrective procedures/recommendations.

In a quasi-experimental design study by Chavez & Mitchell (2020), faculty members teaching identical online courses recorded welcome videos that were presented to students at the course onset, constituting the sole exposure to perceived gender and race/ethnicity. Results showed that instructors who are female and persons of color received lower scores on ordinal student evaluations than those who are white males. Chavez & Mitchell (2020) posits that there are potential direct and indirect biases against instructors of other races and ethnicities. Because gender, race and ethnicity are highly visible, these make them frequently tapped stereotypes. Based on role-incongruity theory and the notion of marginality, Chavez & Mitchell (2020) expect that scholars of color receive systematically lower evaluation scores, thereby stunting their competitiveness with colleagues born more conveniently into students' stereotypes. Just as gendered evaluations operate on "shifting standards" where one making a judgment is compelled to do so relative to a reference point, Chavez & Mitchell (2020) asserts that the same process occurs with people of color and accent in comparison to white males with native linguistic inflections.

In a study, examining the effects of professor's race and clothing style on student evaluations, Aruguete, Slater and Mwalkinda (2017) randomly assigned students to one of four conditions in which they received a photograph and description of a professor. The only difference between conditions was the photograph, which systematically varied race (Black or White) and clothing style (casual or formal). Results showed that both Black and White students rated the Black professor less favorably than the White professor. Interestingly, students trusted the Black professor more when he was pictured in formal clothing compared to casual, while the reverse is true for the White professor. Results of this study support previous research which shows Black professors having a significant disadvantage in the student evaluation process when compared with White professors.

A study was conducted from multiple universities to test the presence of an adverse impact against professors belonging to minority groups (African American, Asian American, Hispanic American and foreign national origin) using official student evaluation of teaching (SET). Wang & Gonzalez (2020) conducted a series of regression analyses to compare SET rating sources and control for course difficulty. Results showed that White American professors receive higher SET ratings than non-White American and foreign professors, implying the presence of bias in SET.

A study by Reid (2010) suggests that faculty of color are evaluated worse than white colleagues, especially Black and Asian professors, with Black men faring particularly poorly. As an example, Smith and Hawkins (2011) showed that Black and other non-White faculty received the lowest mean scores across 26 individual multidimensional evaluation items as well as two global measures of course quality, overall value, and overall teaching ability. Anderson (2010) also suggests that people of color may also be punished more for intersectional stereotype nonconformity.

Language Background

Instructor's level of English language proficiency has also been found to affect student ratings (Bosshardt & Watts, 2001; Finegan & Siegfried, 2000; Ogier 2005). Faculty with accents and Asian last names receive lower ratings in both SETs and Rate My Professor (Fan et al, 2019; Subtirelu, 2015). In addition, faculty with accents fare worse than their white and native English-speaking counterparts (Kreitzer & Sweet-Cushman, 2021). While Saunders (2001) did not find differences in evaluations of instructors whose native language is English

compared to those for whom English is a second language, Gill (1994) found that students view teachers with "standard North American accents" more favorably.

Rank and Tenure

Findings on the impact of student evaluations according to the faculty members' status, rank, and tenure are mixed. While some have found that non-tenured faculty receive lower ratings than tenured faculty (e.g., McPherson & Jewell, 2007), others have found that adjunct and temporary faculty tend to receive higher ratings than tenure-track faculty (Figlio, Schapiro & Soter, 2015; McPherson et al., 2009). There does not appear to be a consistent or systematic difference among the ratings of full professors compared with associate professors or of junior versus senior lecturers (Spooren, 2010; Ting, 2000).

Faculty and Student Perceptions

Research has shown that student evaluations are influenced by whether students perceive the evaluation process as making a difference. Chen and Hoshower (2003) found that students are motivated to participate in student evaluations "by the expectation that they will be able to provide meaningful feedback" (p. 71). Furthermore, Worthington (2002) found that "students who perceive the evaluation process as a process for improving teaching in the future…have a higher probability of giving a more favourable ranking" (p.61).

Other research shows that students may not believe that the opinions they express on their evaluations are taken seriously by faculty or administrators (Spencer & Schmelkin, 2002). Richardson's (2005) comprehensive review of literature on student evaluations concluded that "[m]any students and teachers believe that student feedback is useful and informative, but for a number of reasons many teachers and institutions do not take student feedback sufficiently seriously" (p. 387).

Some studies find that information from student evaluations does not contribute to changes in teaching practices (Blair & Valdez Noel, 2014; Kember et al., 2002; Nasser & Fresko, 2002; Spencer & Flyr, 1992). Others, however, find that student evaluations are generally perceived as useful for "formative and summative" purposes (Schmelkin et al., 1997, p. 588) and may lead to changes in instruction (Beran et al., 2005; Chan et al., 2014; Gravestock & Gregor-Greenleaf, 2008; Panasuk & Lebaron, 1999). Arthur (2009) lists four reasons why faculty might not make changes in response to student evaluations: 1) "the issue was felt by just one student," 2) "students complained about difficult concepts which were nevertheless important for them to learn," 3) "students did not know what would be useful to them in the workplace, so asked for inappropriate changes," and 4) "students' comments seemed to fly in the face of the facts" (p. 450).

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andersen, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *Political Science & Politics, 30*, 216-219.

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles, 49*, 507-516.

Arreola, R.A. (2000). *Developing a comprehensive faculty evaluation system*. Bolton, MA, Anker Publishing.

Arthur, L. (2009). From performativity to professionalism: Lecturers' responses to student feedback. *Teaching in Higher Education, 14*, 441-454.

Aruguete, M. S., Slater, J., & Mwaikinda, S. R. (2017). The effects of professors' race and clothing style on student evaluations. *The Journal of Negro Education*, *86*(4), 494-502.

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *The Journal of Economic Education, 37,* 21-37.

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*, 193-210.

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*, 656.

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*, 91-106.

Basow, S., Codos, S., & Martin, J. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal, 47*, 352-363.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30*, 25-35.

Beran, T. N., & Rokosh, J. L. (2009). Instructors' perspectives on the utility of student ratings of instruction. *Instructional Science, 37*, 171-184.

Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The utility of student ratings of instruction for students, faculty, and administrators: A "Consequential Validity" study. *Canadian Journal of Higher Education, 35*, 49-70.

Blair, E., & Valdez Noel, K. (2014). Improving higher education practice through student evaluation systems: Is the student voice being heard? *Assessment and Evaluation in Higher Education, 39*, 879-894.

Bosshardt, W., & Watts, M. (2001). Comparing student and instructor evaluations of teaching. *The Journal of Economic Education, 32*, 3-17.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41*, 71-88.

Cao, Y., Clark, A., Schrimer, J., & Nelson, M. (2007). *Online and paper course evaluations: Are the response rates and results different?* Paper presented at the Association of Institutional Research Annual Forum, San Francisco, CA.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*, 17-33.

Chamberlin, M. S., & Hickey, J. S. (2001). Student evaluations of faculty performance: The role of gender expectations in differential evaluations. *Educational Research Quarterly, 25*, 3.

Chan, C. K., Luk, L. Y., & Zeng, M. (2014). Teachers' perceptions of student evaluations of teaching. *Educational Research and Evaluation, 20*, 275-289.

Chang, T. S. (2003). *The results of student ratings: The comparison between paper and online surveys.* Paper

presented at the annual meeting of American Educational Research Association, Chicago, IL.
Chapman, L., & Ludlow, L. (2010). Can downsizing college class sizes augment student outcomes? An investigation of the effects of class size on student learning. *The Journal of General Education, 59*, 105-123.
Chavez, L. & Mitchell, K. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity, *PS: Political Science & Politics, 53,* (2) 270-274. doi:10.1017/S1049096519001744
Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education, 28*, 71-88.
Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16-30.
Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education*, *13*(4), 321-341.
Cuseo, J. (2007). The empirical case against large class size: Adverse effects on the teaching, learning, and retention of first-year students. *The Journal of Faculty Development*, 21(1), 5-21.
Cooper, K. M., Downing, V. R., & Brownell, S. E. (2018). The influence of active learning practices on student anxiety in large-enrollment college science classrooms. *International Journal of STEM Education*, *5*(1), 1-18.
Dalmia, S., Giedeman, D. C., Klein, H. A., & Levenburg, N. M. (2005). Women in academia: An analysis of their expectations, performance and pay. *Forum on Public Policy, 1,* 160-177.
Donovan, J., Mader, C. E., & Shinsky, J. (2006). Constructive student feedback: Online vs. traditional course evaluations. *Journal of Interactive Online Learning, 5*, 283- 296.
Downing, V. R., Cooper, K. M., Cala, J. M., Gin, L. E., & Brownell, S. E. (2020). Fear of negative evaluation and student anxiety in community college active-learning science courses. *CBE—Life Sciences Education*, *19*(2), at 20.
Duckor, B., & Stark, P. (2021, March). Evaluating the evaluation of teaching in higher education: What the data from student surveys do and do no not tell us. *BEAR Seminar*, University of California, Berkeley, California.
Duckor, B., Chen, C., Currin-Percival, M., Tortora, C., & Villagran, M. (2024, February). *SOTE instrument technical report*. Student Evaluation Review Board Committee, San José State University. [Technical Report]. https://drive.google.com/file/d/1WWhIZ2lZ-sKiCbTMv89D_e4UNfifJPMM/view?usp=sharing
Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS One,14*(2), e0209749
Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, *29*(1), 91-108.
Figlio, D., Schapiro, M., & Soter, K. (2015). Are tenure track teachers better? *Review of Economics and Statistics, 97*, 715-724.
Finegan, T. A., & Siegfried, J. J. (2000). Are student ratings of teaching effectiveness influenced by instructors' English language proficiency? *The American Economist, 44*, 17-29.
Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111
Gill, M. M. (1994). Accent and stereotypes: Their effect on perceptions of teachers and lecture comprehension. *Journal of Applied Communication Research, 22*, 348-361.
Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends.* Toronto: Higher Education Quality Council of Ontario.
Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209-1217.
Guder, F., & Malliaris, M. (2013). Online course evaluations response rates. *American Journal of Business Education, 6*, 333-337.

Hardy, N. (2003). Online ratings: fact and fiction. *New Directions for Teaching and Learning, 96*, 31-41.
Heath, N. M., Lawyer, S. R., & Rasmussen, E. B. (2007). Web-based versus paper-and-pencil course evaluations. *Teaching of Psychology, 34*, 259-261.
Heffernan, T. (2022). Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education*, *47*(1), 144-154.
Henderson, C., Khan, R., & Dancy, M. (2018). Will my student evaluations decrease if I adopt an active learning instructional strategy?. *American Journal of Physics*, *86*(12), 934-942.
Johnson, M. D., Narayanan, A., & Sawaya, W. J. (2013). Effects of course and instructor characteristics on student evaluation of teaching across a college of engineering. *Journal of Engineering Education, 102*, 289-318.
Kember, D., Jenkins, W., & Chi Ng, K. (2004). Adult students' perceptions of good teaching as a function of their conceptions of learning—Part 2. Implications for the evaluation of teaching. *Studies in Continuing Education*, *26*(1), 81-97.
Kember, D., Leung, D. Y., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education, 27*, 411-425.
Kohn, J., & Hatfield, L. (2006). The role of gender in teaching effectiveness ratings of faculty. *Academy of Educational Leadership Journal, 10*, 121.
Kolitch, E., & Dean, A. V. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about 'good s teaching. *Studies in Higher Education*, *24*(1), 27-42.
Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings?. *Frontiers in psychology*, 570.
Kreitzer, R.J., & Sweet-Cushman, J. (2022). Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform. *J Acad Ethics 20*, 73–84. https://doi.org/10.1007/s10805-021-09400-w
Kulik, J. A. (2001). *Student ratings: Validity, utility, and controversy*. In M. Theall, P. C. Abrami, & Mets, L.A. (Eds), The student rating debate: Are they valid? How can we best use them? (pp. 9-25). San Francisco: Jossey-Bass.
Låg, T., & Sæle, R. G. (2019). Does the flipped classroom improve student learning and satisfaction? A systematic review and meta-analysis. *AERA open*, *5*(3), 2332858419870489.
Laubsch, P. (2006). Online and in-person evaluations: A literature review and exploratory comparison. *Journal of Online Learning and Teaching, 2*, 62-73.
Liu, Y, (2006). A comparison study of online versus traditional student evaluation of instruction. *International Journal of Instructional Technology & Distance Learning.*
MacNell, L., Driscoll, A., & Hunt, A. N. (2014). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*, 291-303.
Mandel, P., & Süssmuth, B. (2011). Size matters. The relevance and Hicksian surplus of preferred college class size. *Economics of Education Review, 30*(5), 1073-1084.
Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7,* 9-18.
Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187.
Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal, 16*(1), 57-70.
Martin, L. L. (2016). Gender, teaching evaluations, and professional success in political science. *Political Science & Politics, 49*, 313-319.
Mateo, M.A., & Fernandez, J. (1996). Incidence of class size on the evaluation of university teaching quality. *Educational and Psychological Measurement, 56*, 771-778.
Mau, R. R., & Opengart, R. (2012). Comparing ratings: In-class (paper) vs. out of class (online) student evaluations. *Higher Education Studies, 2*(3), 55.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, 15(2), 51–69.
McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly, 88*, 868–881.
McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal, 35,* 37–51.
Mengel, F., Sauermann, J. & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association, 17*(2), 535-566.
Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *Political Science & Politics, 51*, 1-5.
Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27*, 187-198.
Ogier, J. (2005). Evaluating the effect of a lecturer's language background on a student rating of teaching form. *Assessment & Evaluation in Higher Education, 30*, 477-488.
Panasuk, R. M., & Lebaron, J. (1999). Student feedback: A tool for improving instruction in graduate education. *Education, 120*, 356-356.
Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors. *Journal of Diversity in Higher Education, 3*(3), 137
Richardson, J. T. E. (2005). Instruments for obtaining feedback: A review of the literature. *Assessment & Evaluation in Higher Education, 30*, 387-415.
Samuel, M. L. (2021). Flipped pedagogy and student evaluations of teaching. *Active Learning in Higher Education*, *22*(2), 159-168.
Saunders, K. T. (2001). The influence of instructor native language on student learning and instructor ratings. *Eastern Economic Journal, 27*, 345-353.
Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education, 38*, 575-592.
Sijtsma, K. (2015). Delimiting Coefficient α from Internal Consistency and Unidimensionality. *Educational Measurement: Issues & Practice,* 34(4), 10–13. https://doi-org.libaccess.sjlibrary.org/10.1111/emip.12099.
Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin, 26*, 1329-1342.
Smith, B. P. (2007). Student ratings of teaching effectiveness: An analysis of end-of-course faculty evaluations. *College Student Journal, 41*, 788-801.
Smith, B. P., & Hawkins, B. (2011). Examining student evaluations of black college faculty: Does race matter? *The Journal of Negro Education,* 149-162.
Smith, B. P., & Johnson-Bailey, J. (2011). Student ratings of teaching effectiveness: Implications for non-White women in the academy. *Negro Educational Review, 62*, 115.
Sorenson, L. & Johnson, T. (2006). *Online Student Ratings of Instruction.* Paper presented at Town Hall Meeting April 12, 2006, San Jose, CA.
Spencer, P. A., & Flyr, M. L. (1992). The Formal Evaluation as an Impetus to Classroom Change: Myth or Reality?
Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education, 27*, 397-409.
Spooner, F., Jordan, L., Algozzine, R., & Spooner, M. (1999) Student rating of instruction in distance learning and on-campus classes. *The Journal of Educational Research, 92*, 132-140.
Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation, 36,* 121–131.
Spooren, P., & Mortelmans, D. (2006). Teacher professionalism and student evaluation of teaching: will better teachers receive higher ratings and will better students give higher ratings? *Educational Studies, 32*(2), 201 - 214.
Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can

hurt us. *Sex Roles, 53*, 779-793.
Stark. P. B. (2018 May-June). *Student Evaluations of Teaching are Not Valid. It is time to stop using SET scores in personnel decisions*, John W. Lawrence, American Association of University Professors. https://www.aaup.org/article/student-evaluations-teaching-are-not-valid
Stark, P. B. (2018 Oct). *Student evaluations of teaching do not measure teaching effectiveness. What do they measur*e?, Stanford-Berkeley Joint Colloquium, Department of Statistics, Stanford University, Stanford, CA. https://www.stat.berkeley.edu/~stark/Seminars/setStanford18.htm
Stark, P. B. (2019). *Notes on Student Evaluations of Teaching (SET)*. https://www.stat.berkeley.edu/~stark/Preprints/setNotes19.pdf
Stark, P. B. (2021 March). *Evaluating the Evaluation of Teaching in Higher Education: What the Data from Student Surveys Do and Don't Tell Us*, Berkeley Evaluation and Assessment Research (BEAR) Center, University of California, Berkeley, CA. https://www.stat.berkeley.edu/~stark/Seminars/setUCBED21.htm
Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. Basic and Applied Social Psychology, 42(4), 276-294.
Subtirelu, N. C. (2015). "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors. com. *Language in Society, 44*(1), 35–62
Sujitparapitaya, S. & Briggs, J. (2010). *Does a Delivery Method Matter? A Comparison between Online and Paper Teaching Evaluations.* Paper presented at the Student Evaluation Review Board, San Jose, CA.
Theall, M. (2010). New resistance to student evaluations of teaching. *The Journal of Faculty Development*, *24*(3), 44.
Theall, Michael, & Franklin, Jennifer. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research, 27*, 45-56.
Ting, K. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*, 637–661.
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.
Van Mol, C. (2017). Improving web survey efficiency: the impact of an extra reminder and reminder content on web survey response. *International Journal of Social Research Methodology, 20*(4), 317-327.
Walker, J. D., Cotner, S. H., Baepler, P. M., & Decker, M. D. (2008). A delicate balance: integrating active learning into a large lecture course. *CBE—Life Sciences Education*, *7*(4), 361-367.
Wang, L., & Gonzalez, J. A. (2020). Racial/ethnic and national origin bias in SET. *International Journal of Organizational Analysis, 28(4),* 843-855.
Worthington, A. C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment & Evaluation in Higher Education, 27*, 49-64.