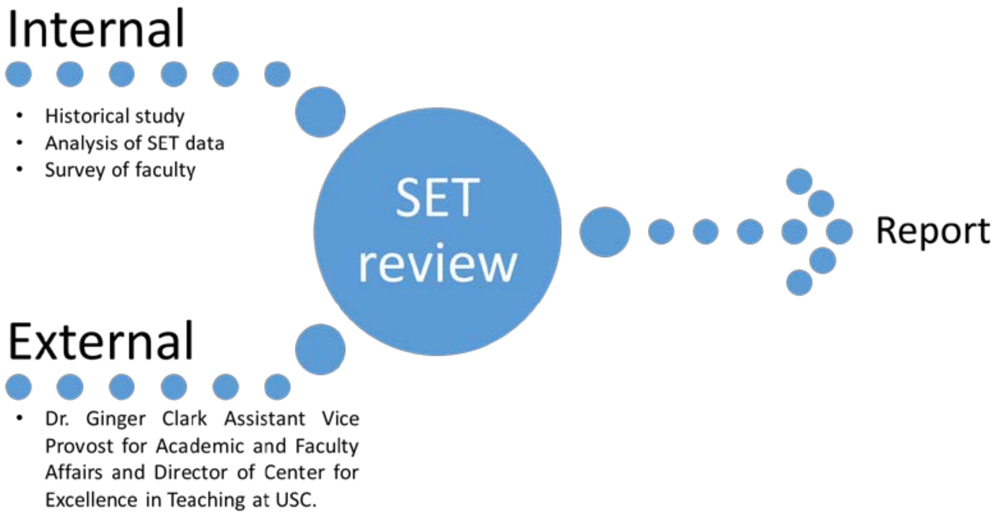
CEPP Report on qSET analysis

Brief Background

The College has long sought student feedback to evaluate teaching at Skidmore. The College, relevant committees, and faculty adopted a 3-question format over two decades ago to gain insight into students' evaluation of teaching effectiveness. In an effort to assist faculty in receiving more nuanced, and less biased, feedback CEPP and ATC (formerly known as CAPT) formed a joint subcommittee in 2011 to evaluate and potentially revise the 3-question student evaluation of teaching instrument. Based on the work of this subcommittee, along with CEPP input, internal and external consultation, and pilot testing, the current "Student Rating of Courses and Teaching" form was born and ultimately adopted by the faculty (See: https://www.skidmore.edu/dof-vpaa/meetings/faculty/2012-2013/minutes3-1.php). The rationale outlined in the motion put forth by CEPP proposed that "an assessment of this new form be conducted no later than the academic year 2016/2017, with that assessment shared with the faculty." Likely due to CEPP's focus on the general education curriculum overhaul, this assessment was postponed until the current academic year. To conduct this analysis of the quantitative student evaluation of teaching (qSET), CEPP devised the following process outlined in the figure below:


Internal
Historical study
Analysis of SET data
Survey of faculty
SET review
Report
External
Dr. Ginger Clark Assistant Vice Provost for Academic and Faculty Affairs and Director of Center for Excellence in Teaching at USC.

Historical Study

To understand the genesis of the current qSET form, CEPP read the historical context and the work that went into developing and testing the form. Michael Arnush, the faculty member who chaired CEPP during the development of the current qSET, met with CEPP in spring 2018. Subsequently, a representative from CEPP (S.Ives, Co-Chair) also met with Catherine Berheide, a faculty member who was directly involved in the refinement and testing of the instrument. In summary, the current qSET was piloted in the 2012 FYE in conjunction with a National Science Foundation award, and was developed and refined in broad consultation with faculty internally, and through seeking input from external experts in the field as well (Joey Sprague and Catherine Ross). In the spirit of following a similar process, we sought to perform an internal analysis of the qSET form and obtain external input.

Internal Assessment of qSET

There a very few topics that bring about such passionate discourse as the topic of student evaluations of teaching, so it is no surprise the topic has been a longstanding and active area of scholarly investigation. Published work has raised concerns about the following: what such forms are actually measuring [1], potential bias against women [2], potential bias against people of color [3], potential bias against non-native English speakers [4], disparity between disciplines, and potential issues in statistical interpretation [5]. Thus, CEPP felt it prudent to conduct an internal analysis of the qSET data over the last 5 years (2013-2018) since the form was adopted.

Analysis of qSET Data (2013-2018)

CEPP partnered with the Office of Institutional Research to gather and analyze the qSET data, as well as data on potential factors that might contribute to or explain some of the scores (e.g. academic division). The specific information obtained, from the form itself and other sources, were: academic division, course level, course start time, reason for taking course, other course characteristics (e.g. enrollment, credits, fall vs. spring), student information (desire to enroll, hrs/week devoted to course, expected grade), course student composition (class year, % of class identifying as ALANA [African, Latinx, Asian, Native American], % of class female, and response rate), instructor (ALANA, international, gender, and age), employment status (fulltime regular, fulltime temporary, part-time regular, part-time temporary). CEPP also sought to determine whether there was agreement between the sub questions in each category and the overall questions. Finally, we were curious if teaching evaluations were changing over time. To summarize, we aimed to determine:

1) if certain student, course, or instructor characteristics influenced qSET scores at Skidmore
2) if there was congruity between sub-question and overall question qSET scores
3) if qSET scores are changing over time

Prior to completing the analysis, the data were first cleaned (e.g. exclude missing data) and some courses were removed for varying reasons. The only courses removed were science labs (these due to potential double-counting, not graded separately, etc.), PA riding courses, and private music instruction courses. The data were then analyzed, using multivariate analysis, to flag whether something was, or was not, statistically significant—meaning if an effect was found it is unlikely to be due to chance alone—and the strength or magnitude of the effect was also recorded. This analytical approach is similar to but more expansive than what the Dean of Faculty and Office of Institutional Research undertook in 2013-2014 and presented to Chairs and Program directors in 2015. In order to maintain anonymity, members of CEPP felt it appropriate to never have access to, see, or directly use the raw data, and this responsibility was delegated to the OIR. Using the data analysis, CEPP, and not OIR, drafted the present report, as part of this broader assessment. Finally, due to logistical issues, the analysis was conducted on "course section level" data, and not "student level," meaning the data entered were for each unique course section per term, or the average of all the students’ data for the course section, and not every single student response which is the way that the data are reported to faculty.

CEPP noted three general characteristics of the data:

- Average qSET scores are generally quite high (~4.5 out of 5), with the bulk of individual ratings falling in the 4-5 range, but the scores still vary by nearly a full point when considering student-level data. The variation in scores within a question is an important consideration, as it is unlikely that small deviations in qSET scores represent significant deviations from the average.
- qSET scores trend upward over time, increasing about 3% from 2013 to 2018.
- Consistent with the pilot study of the current form, there is a very high degree of congruity (correlation ~0.94) between a composite score of sub-questions and the independent “overall” questions for “the course,” for example.

In terms of student, course, and instructor characteristics, several factors were found to be statistically significant (i.e. not likely due to chance alone), but the magnitude of the effects were relatively small compared to the overall variation in scores (see table appended below). The items that had the largest effect were “students desire to enroll in the course” and “expected grade,” both of which were positively related to qSET scores, meaning the greater the desire to enroll in the course and the higher the expected grade, the higher the qSET scores. The increase in qSET scores was about 0.1 and 0.3 on a 5-unit scale for each incremental change in “expected grade” or “desire to enroll in the course,” respectively. To put this into context, if students in a class all reported one unit greater desire to enroll in the course—e.g. “more than other courses” (a 4 on the qSET form) vs. “about the same as other courses” (a 3)—would impact qSET scores by about 5-7%, depending on the year and qSET item (e.g. course overall).

The impact of all other factors included in the analysis appear smaller (<0.1 out of 5), amounting to effects of 2% or less on qSET score for a given change in a factor. For example, instructor age was always found to be a significant negative predictor; that is, the older age, the lower the qSET score. However, the effect of age was -0.005/year, so taking an extreme example, a 65 year old professor would be expected to have a 0.18 lower score on “instructor overall” than a 30 year old professor.

Previous research has demonstrated that a faculty member’s race or ethnicity may negatively impact qSET scores, specifically negatively impacting faculty of color [3]. It is worth noting that in 3 of 5 years, an instructor identifying as ALANA had significantly lower qSET scores, averaging 0.04-0.06 points lower out of 5, or negatively impacted by 0.8 to 1.2% over the 5 years. A prior study by McPherson and Jewell [3] found that faculty of color were rated 0.08 points lower on a 4 point scale or 2% lower. This negative bias against faculty of color found in that prior study remained after ruling out potential influence of other measureable factors (e.g. years teaching). Thus, while bias at Skidmore might be of less magnitude than published work, this negative bias is not zero.

Interestingly, the instructor’s gender does not significantly impact the results. If anything, females consistently score on average 0.01 higher on qSET scores. This finding differs from some published studies that suggest females can be disadvantaged by nearly half a point [2].

It is worth noting that the above data analysis was conducted in an exploratory fashion. Thus, we did not adjust our criteria for significance in accordance with the number of questions or variables being explored (e.g. does time of day influence qSET scores). Specifically, based upon probability, the more questions we ask or variables we explore, the more likely we might find a “significant” result simply because we asked many questions or explored many variables. In statistics, many would suggest that researchers adjust their level of significance, or alpha level, to compensate for the number of questions or variables being probed [6]. Such an adjustment to what was deemed statistically significant, either a priori or post hoc, was not made in the current analysis.

The level of the course does appear to influence qSET scores, since scores appear to fall as the course level increases. 200- and 300-level courses are significantly scored lower than 100-level courses.

There are some additional factors that were looked at. For example, instructor status (e.g. instructor, tenure-track, or tenured) was never significant. Another item for consideration was expected grade vs. actual grade and the effects and trends were similar. Interestingly, neither expected nor actual grades differed over time.

Though most effects were a fraction of a point and some were statistically insignificant, it may still be that an individual instructor’s scores are impacted by the sum of the various factors. For example, an older professor teaching a required course (low desire to enroll) at an upper level may see lower scores as a result of the combined effect of these course and instructor characteristics.

As is often the case with research, a project can bring about more questions than it answers, and the robust analysis conducted in collaboration with the Office of Institutional Research is not meant to either exonerate or condemn the qSET necessarily, but certainly provides information deserving of consideration by key parties (ATC, PC, etc.) and the faculty as a whole. The outputs of the analysis will be made fully available to the Skidmore community, though to protect confidentiality of individual instructors scores, the raw data will not be made publicly available.

Analysis of Faculty Knowledge and Perceptions of the qSET form

In accordance with guidance set forth in the CEPP operating code, which states that CEPP will “ensure extensive, widespread and high quality consultation take place during all major initiatives,” we devised and conducted a survey of all faculty to gain insight into faculty perceptions of the Skidmore qSET and qSETs more broadly. This mechanism was designed to be more inclusive and anonymous in nature, and gain feedback from a larger number of faculty than typically engage in faculty floor or open forum discussions.

The survey was made available to the faculty email list. From this, 168 faculty completed the survey, with 48 non-tenure-track (TT), 39 tenure-track, and 80 tenured faculty. Approximately 40% were current or former department chair/program directors and about 15% had served on the tenure and promotion committee. Collectively, the sample was relatively robust and representative of the College.

The main findings of the survey are:

- In terms of satisfaction with the current qSET form, and whether the form provides useful feedback, faculty were divided and well dispersed from strongly disagree to strongly agree. 53% disagreed or disagreed strongly that they are satisfied with the qSET form.
- 66% of faculty disagreed or disagreed strongly that the qSET are objective evaluations of teaching.
- 77% of faculty disagreed or disagreed strongly that instructor characteristics do not influence qSET scores.
- 69% of faculty disagreed or disagreed strongly that qSET is useful to assess student learning, and 59% disagreed or disagreed strongly that students base their ratings on how much they learned. 62% disagreed or disagreed strongly that faculty qSET scores correlate with student learning outcomes. 53% disagreed or disagreed strongly that qSET scores correlate with future academic performance.
- While 50% of faculty disagreed or disagreed strongly that the qSET data are necessary for evaluating teaching, about 23% agreed or agreed strongly.
- 61% of faculty disagreed or disagreed strongly that there is a specific qSET score that demonstrates a minimum standard of effective teaching. Of those that agreed, the responses ranged from 2 to 3 to 3.5 to 4. Others mentioned the mean or median as minimum scores. One response stated “4.0 as the "gold standard,” or “People seem to think that it's 4. I don't share that belief,” and another that “This is absolutely where bias could come in.”
- 88% of faculty disagreed that a qSET score less than the mean would indicate inadequate teaching. Of those that disagreed, 4% said qSET scores more 1 standard deviation (SD) would be ineffective, 12% said more than 2 SD, and 84% said unclear/need more information.
- Faculty were divided and well dispersed from strongly disagree to strongly agree as to whether students give higher ratings to those who teach less demanding courses, with TT faculty more likely to agree than other cohorts.
- 80% agreed or agreed strongly that students base their ratings on satisfaction with the course or instructor.
- Faculty were divided and well dispersed from strongly disagree to strongly agree as to whether they as an instructor feel pressure to make their course less demanding. TT faculty were less likely to disagree than tenured or NTT faculty.
- 77% and 73% of faculty agreed or agreed strongly that faculty gender and race/ethnicity, respectively, influences their qSET scores.
- 59% of faculty agreed or agreed strongly that instructors take the qSET seriously. However, there was no consensus whether students take them seriously, though 62% of faculty believed students should be trained about the qSET.
- 56% of faculty disagreed or disagreed strongly that they are satisfied with how the qSET is used.

- 80% of faculty agreed or agreed strongly that more comprehensive methods should be used to evaluate teaching effectiveness. 74% of faculty agreed or agreed strongly that too much weight is placed on qSET.
- The open-ended comments ranged significantly.

The results of the faculty survey will be made fully available to the Skidmore community, though some comments may be partially or fully redacted to protect the confidentiality of some individuals and their comments.

External

CEPP and the Dean of Faculty's office invited Dr. Ginger Clark, Professor of Clinical Education in Psychology, Assistant Vice Provost for Academic and Faculty Affairs, and Director of the Center for Excellence in Teaching at the University of Southern California (USC) to campus to speak with specific parties (e.g. teaching support network, CEPP, ATC/PC) and the faculty as a whole. Dr. Clark was chosen as USC recently reviewed and modified its student evaluation of teaching as well as the policies and practices around how teaching effectiveness is evaluated at that institution. In essence, USC is viewing the SET as an indicator of student engagement rather than student learning or teaching effectiveness. A robust peer observation and review process has since been implemented at USC to offset the potential issues surrounding SETs. Dr. Clark shared these experiences and met with specific parties (ATC, PC, CEPP, and the teaching support network fellows), in addition to providing a presentation to the faculty at large.

Conclusions

Significant work by members of the Skidmore community, including broad internal and external consultation over a couple years, went into developing the current qSET form. Since adopting the current form, this report is the first publicly available attempt to review the form, as mandated in the original motion that was put forth and successfully voted on. CEPP finds that there are significant factors influencing qSET scores that are both student- and/or faculty-centric. Coupled with this analysis, faculty were surveyed, and there is a consensus of dissatisfaction with the current form and how it is used, and faculty support the idea that comprehensive methods of assessing teaching effectiveness should be used to better balance use of the qSET. Based upon the outcomes above, CEPP will facilitate further conversations with faculty and will collaborate with relevant committees (e.g. ATC, PC, FEC) to develop strategies to address concerns of the faculty surrounding student evaluations of teaching.

Questions? Any questions regarding this report should be directed to CEPP.

References

1. Uttl, B., C.A. White, and D.W. Gonzalez, *Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related.* Studies in Educational Evaluation, 2017. **54**: p. 22-42.
2. Mitchell, K.M.W. and J. Martin, *Gender Bias in Student Evaluations.* PS: Political Science & Politics, 2018. **51**(03): p. 648-652.
3. McPherson, M.A. and R.T. Jewell, *Leveling the Playing Field: Should Student Evaluation Scores be Adjusted?*.* Social Science Quarterly, 2007. **88**(3): p. 868-881.
4. Subtirelu, N.C., *"She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com.* Language in Society, 2015. **44**(1): p. 35-62.
5. Stark, P.B., Freishtat, R>, *An evaluation of course evaluations.* ScienceOpen Research, 2014. **0**(0): p. 1-7.
6. Chen, S.-Y., Z. Feng, and X. Yi, *A general introduction to adjustment for multiple comparisons.* Journal of thoracic disease, 2017. **9**(6): p. 1725-1729.

Table Summarizing 2013-2018 qSET data analysis

	Course Overall			Instructor Overall			Learning Overall		
		# of years			# of years			# of years	
Independent Variable	Effect on Score?	significant	Effect (%)	Effect on Score?	significant	Effect (%)	Effect on Score?	significant	Effect (%)
Academic Division (relative to Humanities)									
Physical & Life Sciences	-0.05	3	-1.1	-0.06	4	-1.4	-0.03	1	-0.7
Social Sciences	-0.01	0	-0.3	-0.03	1	-0.6	-0.01	0	-0.1
Visual & Performing Arts	-0.05	3	-1.0	-0.06	4	-1.4	-0.04	2	-0.9
Other	-0.06	5	-1.3	-0.07	4	-1.5	-0.04	1	-1.0
Course Level (relative to 100-Level)									
200-Level Courses	-0.04	3	-0.8	-0.03	1	-0.6	-0.03	2	-0.7
300-Level Courses	-0.06	5	-1.4	-0.04	1	-0.8	-0.07	5	-1.5
Start Time (relative to Earliest morning <9am)									
Early morning (9am-9:45am)	0.00	0	0.0	0.00	1	0.1	0.00	0	0.0
Late morning (10am-11:55am)	0.01	0	0.1	0.00	0	0.1	0.01	0	0.2
Early afternoon (12pm-2:50pm)	-0.01	1	-0.2	0.00	1	0.0	0.00	1	-0.1
Late afternoon (3pm-4:40pm)	-0.01	0	-0.2	-0.01	0	-0.2	0.00	0	-0.1
Evening (5pm-8:30pm)	-0.01	1	-0.2	-0.01	2	-0.2	-0.01	1	-0.2
Reason Taking Course									
Required for major	0.00	0	0.0	0.00	0	0.0	0.00	0	0.0
Elective for major	0.00	0	0.0	0.00	0	0.0	0.00	0	0.0
Non-major elective	0.00	0	0.0	0.00	0	0.0	0.00	0	0.0
All-college requirement	0.00	1	0.0	0.00	2	0.0	0.00	1	0.0
Other requirement	0.00	0	0.0	0.00	0	0.0	0.00	0	0.0
Other Course Characteristics									
Course Enrollment #	0.00	4	-0.1	0.00	2	-0.1	0.00	5	-0.1
Course Credits #	-0.02	2	-0.5	-0.03	1	-0.7	-0.02	2	-0.3
Fall Course	-0.01	1	-0.2	-0.01	0	-0.1	-0.01	2	-0.2
Student Information									
Desire to enroll in course	0.30	5	6.7	0.25	5	5.5	0.33	5	7.5
Hours/week devoted to course	0.02	5	0.5	0.01	2	0.3	0.03	5	0.8
Expected grade (0.00-4.00)	0.20	5	4.4	0.20	5	4.4	0.15	4	3.3
Course Composition									
Class Year (1-4 avg)	-0.01	0	-0.2	-0.01	0	-0.2	0.01	1	0.3
% Class ALANA	0.00	1	0.0	0.00	1	0.0	0.00	2	0.0
% Class Female	0.00	3	0.0	0.00	3	0.0	0.00	3	0.0
Response Rate	0.00	2	0.0	0.00	2	0.0	0.00	2	0.0
Instructor									
Instructor ALANA	-0.05	3	-1.1	-0.06	3	-1.3	-0.04	3	-1.0
Instructor International	-0.03	1	-0.6	-0.03	0	-0.7	-0.02	0	-0.4
Instructor Female	0.01	0	0.2	0.01	0	0.3	0.01	0	0.2
Instructor Age	0.00	5	-0.1	-0.01	5	-0.1	0.00	4	-0.1
Employment Category (relative to Fulltime-Regular)									
Fulltime-Temporary	-0.04	2	-0.8	-0.04	2	-1.0	-0.04	3	-0.8
Parttime-Regular or Shared Parttime-Regular	0.01	0	0.3	0.01	0	0.2	0.02	0	0.4
Parttime-Temporary	-0.09	5	-2.1	-0.10	5	-2.3	-0.08	5	-1.8