Fresno State Student Ratings of Instruction (FSSRI) Questionnaire:

Report on Scale Construction, Reliability, and Validity

Kathleen Dyer, on behalf of the

Academic Senate Task Force on Fresno State Student Ratings Questionnaire

California State University, Fresno

September 9, 2019

Fresno State Student Ratings of Instruction (FSSRI) Questionnaire:

Report on Scale Construction, Reliability, and Validity

Background

In Spring semester of 2018, with one year remaining of a contract with IDEA for student ratings, the Fresno State Academic Senate voted to explore options in addition to continuing with IDEA's soon-to-be fully online course evaluation program. The reasons included:

- IDEA was transitioning to a fully online system. Not all faculty at Fresno State were comfortable moving to online evaluations without the option of conducting evaluations on paper. And the shift to online evaluation included a big price increase.
- Some dissatisfaction with IDEA was identified among the faculty. Complaints included:
    - The system is opaque. It was unclear how they calculated their "raw" and "weighted" scores. Those who requested clarification were directed to white papers that were equally opaque.
    - It appears that mandatory items (e.g., "overall, this is an excellent course") comprise a large portion of the summary scores. This does not conform to the requirement in the APM that we choose from a pool of approved items.
    - Low scores were difficult to act upon because they don't directly address what teachers do in the classroom.
    - The APM requires us to evaluate specific dimensions of teaching. Those dimensions are not addressed at all in the IDEA rating system. Therefore, it seemed badly out of compliance with our own policy.
    - IDEA is based on the premise that students can self-evaluate how much progress they had made on learning outcomes. Decades of research demonstrates clearly that human beings are inaccurate reporters of their own knowledge and skill, known as the "overconfidence effect" (e.g., Svenson, 1981; Kruger & Dunning, 1999; Atir, Rosenzweig, & Dunning, 2015; Kardas & O'Brien, 2018).

The Senate formed a Task Force to explore the possibility of creating a Fresno State Student Ratings of Instruction questionnaire. That Task Force was formed by the end of the Spring semester, with representation from all eight colleges on campus. They met in late spring to elect a chair (Kathleen Dyer) and to create a plan for how to approach their charge. A few task force members met over summer 2018 to begin reviewing the existing published literature that would guide their work. The chair of the task force also met with those on campus who had been involved in student ratings or otherwise had some investment in the student rating system. A partial list of those consulted include:

- Director of the Office of Institutional Effectiveness (Xuanning Fu),
- staff member who coordinated evaluations prior to our move to IDEA (Maria Fernandez),
- the AVP of Faculty Affairs (Rudy Sanchez)
- Chair of the Personnel Committee who has worked to revise the policy on evaluation of student ratings (Brian Tsukimura)
- staff member who coordinated online evaluations in the IDEA system (JoLynne Blake)
- staff member who works with Qualtrics and so might be able to create an online system for us (Chris Hernandez)
- representatives of Technology Services and Academic Technology (Robert Guinn & Brent Aurenheimer)
- Director of the Center for Faculty Excellence (Bryan Berrett)

In August 2018, members of the task force visited the faculty assemblies of all colleges but two. In the subsequent weeks, task force members visited meetings of the chairs for those two colleges, as well as a meeting of the California Faculty Association. We provided information about our task and asked for input in the most general way. All of that background gave us the context we needed to consider a reasonable proposal.

Scale Construction

In Fall semester 2018, the task force met weekly, during which time we generated a structure for a new ratings instrument that would meet the requirements of the APM, and identified specific items to be included.

The APM 322 (Policy on the Assessment of Teaching Effectiveness) requires that student ratings address three elements of instruction:

1. Instructional Design, which includes the learning objectives, the syllabus, materials, and organization.
2. Instructional Delivery, which includes strategies and management of the classroom setting.
3. Assessment, which includes measurement of student learning and feedback to students about their performance.

The APM requires that the instrument consist of a pool of items (approved by the Academic Senate and Provost) from which departments may choose items, so that the instrument can be customized. The APM also requires that the instrument have demonstrated reliability and validity.

In order to simultaneously meet all of these requirements, we settled on the strategy of creating multiple small pools of items. That way, choice of items is available, but each pool will be small enough that the reliability of the scale can be demonstrated with and without the optional items.

As for the nature of the items to be included, we established the following principles. We would ask about:

- Directly observable behaviors that are reasonably objective and that students could report on based on their direct experience. We wanted to exclude items asking about how knowledgeable or caring the instructor is, for instance, because that requires students to make inferences and judgements rather than to simply report what happens.

- Instructor behaviors that have been established, in published empirical research, to enhance learning. In other words, the instrument will ask only about evidence-based teaching best practices. We identified the following practices as evidence-based practices in each of the following areas:

  Instructional Design:

  1. Learning objectives are clear in advance so that students know the goal of their studies.
  2. The syllabus is accurate and provides enough information for students to understand expectations.
  3. Course materials are aligned with course content.
  4. Organization of the course is logical and clear.

  Instructional Delivery:

  5. Scaffolding is used, such that new information is connected to prior knowledge.
  6. Active learning strategies are used, such that students help to construct knowledge.
  7. Connections are made between course content and students' lives, so that they can see how the course is meaningful.
  8. Environment is welcoming and sensitive to student needs.

  Assessment:

  9. Frequent, lower-stakes assessments are used so that improvement is possible over the semester
  10. Grading is timely such that students have adequate information about how they are doing
  11. Instructions are clear as to the purpose of graded assignments
  12. Feedback is offered that is specific and constructive

Furthermore, we decided that since we would not have enough time to construct items from scratch and field test them, we would use items that are already in use elsewhere. We identified several item banks made available by other universities and some that are provided in published research on student ratings.

Therefore, we combed through existing item banks to find items that fit into each of the 12 evidence-based practices described above. We occasionally made slight revisions to the existing items, but only within specific parameters. Specifically, some items were double-barreled, and so we separated them into two items. For instance: "Graded exams and assignments were returned in a timely fashion" was converted into two items: "Graded exams were returned in a timely fashion" and "Graded assignments were returned in a timely fashion." For the sake of consistency, we changed items so that they all referred to the "instructor" rather than the "teacher" or "professor," and we put all items in the past tense. Finally, we duplicated some items and revised one version slightly to refer specifically to labs and to off-campus site placements. For instance, we added an item that said "Graded lab reports were returned in a timely fashion."

We finished this process with a list of 66 items organized into 12 pools. From among each pool of items, we selected one to serve as the default.

We could not ask student respondents to respond to all 66 items, so we created various versions of the student ratings instrument. Each version included all 12 default items, plus 12 of the optional items.

Data Collection

During Spring 2019, we asked faculty (via the faculty listserv and by recruitment efforts of task force members) to participate in the pilot study of the new Fresno State Student Ratings of Instruction (FSSRI) questionnaire. Participation included 53 faculty members, 81 course sections, and 2013 completed student surveys.

	Participating Instructors	Participating Course Sections	Participating Students
Jordan College	10	15	374
College of Social Sciences	5	7	157
College of Arts and Humanities	6	9	175
College of Health and Human Services	6	9	219
Lyles College of Engineering	5	7	168
Craig School of Business	5	9	309
Kremen School of Education	5	9	183
College of Science and Math	11	16	427
Total	53	81	2013

The courses were also evaluated using the IDEA instrument, and instructors gave us permission to access their IDEA reports for these courses for the sake of comparison.

Participating faculty members represented all ranks: 18 classes were taught by Professors, 10 by Associate Professors, 47 by Assistant Professors, 3 by full-time lecturers, and 3 by part-time lecturers. Of the 81 classes, 35 had a female instructor, and 46 had a male professor. Instructors self-identified their race/ethnicity: 51 identified as White, 4 as Black, 5 as Asian, 6 as Hispanic, 2 as Middle-Eastern, 3 as White and Middle Eastern, 8 as White and Hispanic, and 2 as other.

Participating courses included lower division (n=16), upper-division (n=57), and credential or graduate(n=8). Most participating classes were taught face-to-face, but 6 were fully online. Participating classes varied in size: 30 were small (< 26 students), 44 were medium (26-50 students) and 7 were large (> 51 students). Some (n=8) included labs, and some (n=5) included a site placement (e.g., internship, practicum, or service-learning).

Results of the analysis of these pilot data are reported below, with statistically significant results highlighted in yellow.

Results – Internal Reliability

Internal reliability is the degree to which items on the scale measure aspects of the same underlying construct. It is assessed statistically using Cronbach's alpha by standards that are widely agreed upon. An alpha of .70 is considered adequate, .80 is good, and .90 is excellent (Cortina, 1993).

The four default items on the Instructional Design subscale have an alpha of = .82. The four default items on the instructional delivery subscale have an alpha of .83. The four default items on the assessment subscale have an alpha of .76. The 12 default items altogether have an alpha of .91. Therefore, we feel very confident that these items are internally reliable, that they have consistency and measure the same underlying construct, which we believe to be teaching quality.

Next, we explored the reliability of the optional items to determine if they can be used interchangeably with the default items.

The 12 lab items have a Cronbach's alpha of .94. Together with the 12 default items, the alpha is .93. The 10 site placement items have a Cronbach's alpha of .94. Together with the 12 default items, the alpha remains .94. The default items combined with the first 12 optional items (Form A) has an alpha of .96. The default items combined 12 other optional items (Form B) has an alpha of .97. The default items combined with the final 12 optional items (Form C) has an alpha of .96. It is clear that the optional items are reliably interchangeable with the default items.

Our conclusion is that the default instrument has excellent internal reliability, and all of the optional items work equally reliably. Therefore, exchanging one item for another does not detract from the internal reliability of the instrument.

Face Validity

We established face validity, a subjective assessment that the items appear to be appropriate, during the process of scale construction. Task force members reviewed and discussed all items individually, each task force member representing the interests of those in their own colleges. We also presented the items to the Academic Senate, and made slight revisions based on feedback received by that faculty body.

Construct Validity

Construct validity is the extent to which the instrument matches the theoretical construct under investigation. In this case, the theoretical construct is our institution's definition of the dimensions of teaching that are outlined in APM 322. The process described above (under "Scale Construction") makes clear that the FSSRI is explicitly very closely tied to the construct of teaching quality, as defined in the APM.

We note here that, although APM 322 describes three dimensions to be included in the student ratings, these data suggest that those dimensions are not meaningfully distinguishable from one another. An exploratory factor analysis of the default items indicated that there is really only one factor to the FSSRI.

Therefore, the subscores should not be used in consideration of personnel decisions because they do not have adequate scientific foundation. Only the total score should be used for any purpose other than providing the instructor with insights about how students perceive the class.

Results – Convergent Validity

We investigated whether the scores on the FSSRI converges in expected ways with three other instruments. We also looked for potential consistency between scores on the FSSRI and student subjective ratings of other aspects of the course.

Post-Secondary Instructional Practices Survey (PIPS)

The Post-Secondary Instructional Practices Survey (PIPS) is a previously validated instrument that characterizes strategies used in college classrooms (Walter, Henderson, Beach, & Williams, 2016). It can be used to differentiate student-centered practices, which are associate with effective learning, from instructor-centered practices, which are not.

We predicted that scores on the FSSRI would correlate positively with student-centered practices, and correlate negatively with instructor-centered practices, as measured by the PIPS instrument. In fact, this is very close to what we found. The total score is positively correlated with student-centered practices on the PIPS (r=.25, p=.03) and not significantly correlated with instructor-centered practices (r=-0.19, p=.09), although it trends in the negative direction.

This is very strong evidence of convergent validity of our new instrument.

Furthermore, we note that the student-centered practice subscore the PIPS is NOT significantly correlated with the IDEA summary score (r=.22, p=.09), suggesting that our new instrument is *more* valid than the one we are replacing.

Instructor-Report of Items

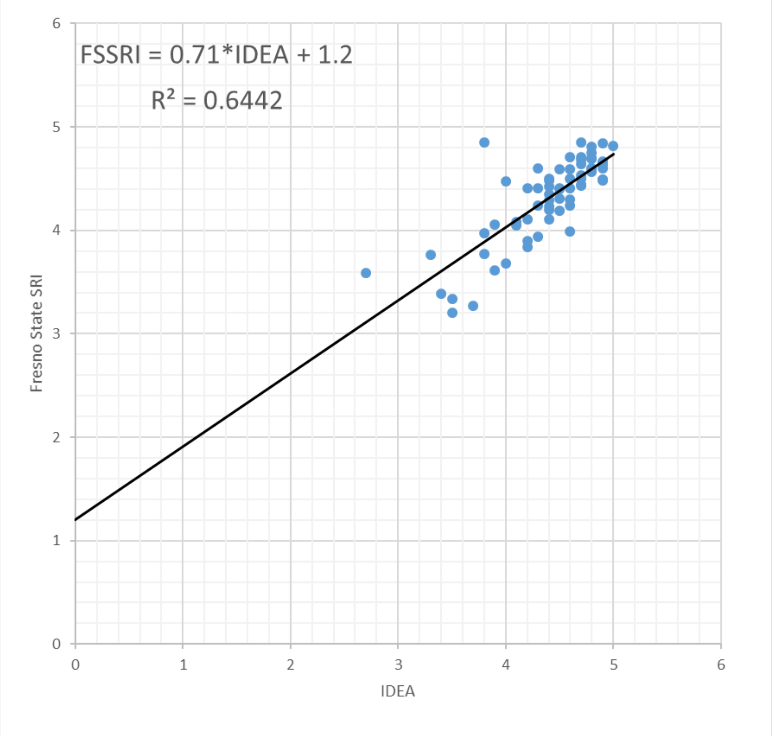
We asked the instructor to self-asses their own teaching strategies in the class using the 12 default items from the FSSRI to see if student-report would converge with instructor-report.

Overall, there is a statistically significant correlation between the total score provided by the instructor and that computed from student surveys (r=0.29, p=.01). This is additional evidence of the validity of the FSSRI instrument.

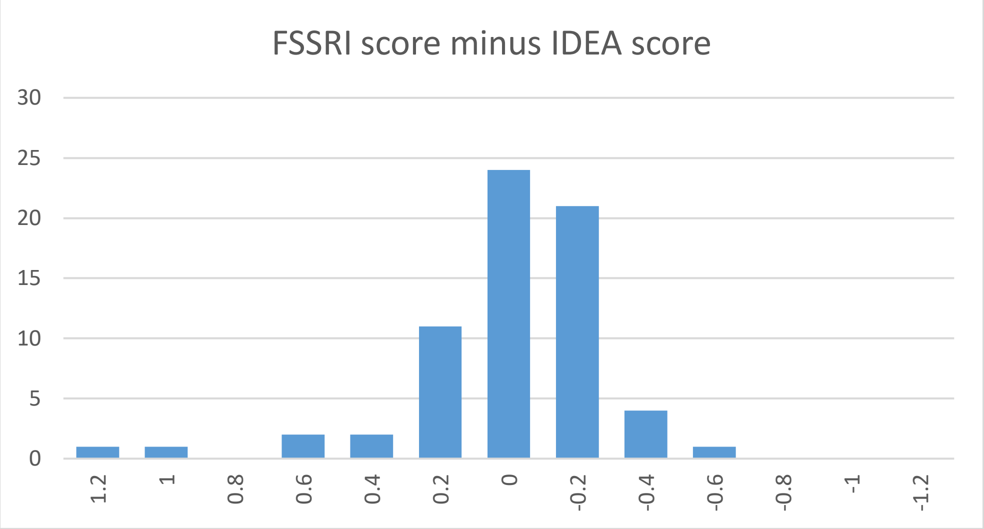
IDEA Ratings

Since the IDEA student rating instrument is based on a fundamentally different principle, we were unsure whether to expect a strong correlation between the new FSSRI scores and the IDEA ratings. However, in the 67 classes for which both scores were available, we found a very strong correlation between to the two instruments: r= .80, p<.0001.

As the following scatterplot demonstrates, both IDEA and FSSRI ratings are clustered heavily at the far upper end of the scale (most scores are higher than 4.0 on a 5-point scale), and FSSRI scores can be predicted reliably by the significant regression line. All of the cases cluster quite close to the regression line.



When the IDEA score is subtracted from the FSSRI score, it is obvious that differences are quite small for most classes. The mean difference between scores is -0.09 (sd=0.28). Therefore, on average, scores with the new instrument are approximately one tenth of a point lower than the IDEA rating for the same class.



Out of the 67 classes in the pilot study with both IDEA and FSSRI data, 84% had a score that was within 0.02 of a point of the IDEA score. 24 had identical scores, 21 had a FSSRI score that was lower by 0.2 of a point, and 11 had a FSSRI score that was higher by 0.2 of a point. Only 2 classes (3%) had a score that differed by 1 full point (on the 5-point scale) and those both scored higher on FSSRI than on IDEA.

Therefore, most faculty will not see much change in their student ratings. As the histogram of differences between the two scores illustrates, two classes in the pilot had an FSSRI score that was a full point (on a 5-point scale) higher than the IDEA rating, but the rest clustered close to the IDEA score.

Other Student Reports

Scores on the FSSRI are also correlated with other student reports. Most notably, the student subjective assessment of the course is strongly correlated with the total score. Correlations were all significant in the expected directions.

	Correlation with total FSSRI score r (p-value)
Student attendance	.07 (p=.001)
Expected grade	.28 (p<.0001)
Frequency of instructor absence	-.12 (p<.0001)
Instructor starts and ends on time	.35 (p<.0001)
Class difficulty	-.22 (p<.0001)
Overall this class was…	.76 (p<.0001)

This is evidence of validity because the instrument is correlated in theoretically expected ways with other factors.

Results – Adjustment of Scores

APM 322 exhorts that “care should be taken to avoid bias” based on protected aspects of identity of the instructor, including race and sex. These are areas of concern because published empirical evidence suggests that there are known biases of students against women and people of color that are elicited in student rating instruments (e.g., MacNell, Driscoll, & Hunt, 2014).

Furthermore, previous research has identified other aspects of the class that are independent of the quality of instruction, such as class size and level, but are known to bias student ratings. Since the faculty member should not be held accountable in their teaching evaluation for such factors, APM 322 directs that “When possible, the instructor should also receive adjusted scores that take into account external factors beyond the control of the instructor.”

Previous research suggests that multiple external factors do influence ratings, but that effect sizes are very small (Beran & Violato, 2005). Such analyses are generally done with samples that include hundreds of thousands of class evaluations, so ours is a pitifully small sample in comparison. Nonetheless, we measured and explored several external factors that have previously been identified as sources of bias, just to see if there are trends.

Course Variables

We looked at class size two ways. First, a simple correlation between number of students enrolled and summary rating was not statistically significant (r = -0.13, p=.26).

Dividing the classes into small (<26 students; n=30), medium (26-50 students; n=44) and large (>50; n=7) also did not demonstrate any statistically significant differences (F=1.24, df=2, p=.30), although there is a trend toward lower ratings for the largest classes.

	FSSRI Total Score Mean (sd)
Small (=n=30)	4.3 (.4)
Medium (n=44)	4.4 (.4)
Large (n=7)	4.1 (.4)

When course level is coded as lower-division, upper-division, and credential/graduate, there is no significant correlation between level and student rating (r=.19, p=.08), although there is a trend toward higher student ratings for credential and graduate level classes.

	FSSRI Total Score Mean (sd)
Lower-Division (=n=14)	4.2 (.4)

Upper-division (n=57)	4.3 (.4)
Post-Bac (n=8)	4.5 (.5)

Instructor Variables

Rank of instructor was not systematically related to summary scores (F=.90, df=4, p=.47).

	FSSRI Total Score Mean (sd)
Part-Time Lecturer (n=3)	4.4 (.1)
Full-Time Lecturer (n=3)	4.1 (.2)
Assistant Prof (n=47)	4.3 (.4)
Associate Prof (n=10)	4.5 (.3)
Professor (n=16)	4.3 (.4)

In our sample, there was no difference between student ratings in classes taught by women as compared to men (F=1.70, df=1, p=.20).

	FSSRI Total Score Mean (sd)
Female Instructor (n=35)	4.4 (.4)
Male Instructor (n=44)	4.3 (.5)

Next, we looked at the self-reported race/ethnicity of the instructor. Most individual racial/ethnic identities were endorsed by so few faculty that it is impossible to analyze them separately. There is no significant difference between those classes taught by White instructors as compared to those taught by non-White instructors (F=1.30, df=1, p=.26).

	FSSRI Total Score Mean (sd)
White Instructor (n=35)	4.3 (.4)
Non-White Instructor (n=19)	4.2 (.5)

Another non-instructional factor on which there may be bias in student ratings is the subject matter of the class. Prior research has found that quantitative subjects are rated more poorly than non-quantitative subjects (Uttl & Smibert, 2017), but this is usually studied by comparing classes in the English department to classes in the Math department. In reality, "quantitative" is more complicated than just these two departments. We examined three different operational definitions of "quantitative": by college, by instructor-report of quantitative content, and by review of the course descriptions and syllabi.

There are statistically significant differences in average scores across colleges (F=3.08,df=7, p=.007). From the highest to the lowest scores, the different is .7 point (on a 5-point scale).

College	FSSRI Total Score Mean (sd)
Lyles (n=7)	3.9 (.4)
Craig (n=9)	4.2 (.5)
CSM (n=14)	4.2 (.4)
CAH (n=9)	4.3 (.5)
COSS (n=7)	4.4 (.3)
Jordan (15)	4.4 (.3)
CHHS (n=9)	4.5 (.3)
Kremen (n=9)	4.6 (.2)

FSSRI scores in colleges whose subjects are primarily quantitative (Lyles, Craig, CSM) are lower than scores in colleges whose subjects are not as likely to be quantitative (CAH, COSS, Jordan, CHHS, Kremen)(F=17.78, df=1, p<.0001) .

Instructors were asked to self-report whether their class was entirely quantitative, partially quantitative, or not at all quantitative. This form of measurement also shows the same pattern, (F=3.51, df=2, p=.04) but the differences between groups are much smaller, only .3 point on a 5 point scale. We suspect that instructors differ on their standards for what they consider “quantitative”.

Quantitative (by instructor report)	FSSRI Total Score Mean (sd)
Not at all (n=39)	4.4 (.4)
Partially (n=29)	4.3 (.4)
Entirely (n=13)	4.1 (.4)

In search of an better operational definition of “quantitative”, we decided that in the more quantitative colleges (Lyles, Craig, CSM) every class would be considered quantitative unless it was described as “conceptual”; in the less quantitative colleges (CAH, COSS, Jordan, CHHS, Kremen) classes would only be considered quantitative if they were about research methods and statistics.

Quantitative (by college and course title)	FSSRI Total Score Mean (sd)
No (n=47)	4.4 (.4)
Yes (n=34)	4.1 (.4)

By this definition, there is a statistically significant difference between the quantitative and non-quantitative courses (F=12.2, df=1, p=.001), with an average difference of .3 points on a 5 point scale.

But it is not clear what is the appropriate operational definition of a “quantitative class”.

Student Variables

We explored variables related to the student, rater than the class or instructor, to see if those were related to ratings.

There is not a significant correlation between student level (e.g., freshman, sophomore, etc.) and student rating (r=-0.3, p=.26).

Ratings differ by the student’s reason for taking the class (F=4.79, df=4, p=.001). It appears that the vast majority of students reported on classes that were required for their degrees, either as core requirements or major electives. General electives were rated about the same as those classes, but GE classes were rated a bit lower, and classes taken for some other (unstated) reason were rated much higher.

Reason for taking class (student report)	Summary Score Mean (sd)
Core course required for my degree (n=1237)	4.3 (.7)
Major elective for my degree (n=346)	4.3 (.7)
General Education (n=253)	4.2 (.7)
Elective (n=80)	4.3 (.7)
Something else (n=64)	4.6 (.5)

Student ratings have a small but statistically significant correlation with student report of their own attendance in class (r=.07, p=.001). There are stronger (also significant) correlations with anticipated grade (r=.28, p<.0001) and perceived difficulty of the class (r=-.22, p<.0001).

While these variables are significantly related to student ratings, it is impossible to identify these as independent of the quality of instruction.

Recommendations

Based on these pilot data, we recommend that:

- All of the tested items be allowed as options.
- Total scores, but not the three subscores, be used for personnel decisions.
- Existing probationary plans can be left unaltered, as there are no drastic differences between the new instrument and IDEA that would alter the establishment of departmental standards.
- No adjustments based on non-instructional factors are recommended at this time. Once the new instrument is adopted, we will continue to monitor institutional data to continue to explore whether adjustments are justified.
- Individual departments may consider revising their policies to lower acceptable standards for quantitative courses. Members of personnel committees and administrators should consider that ratings in quantitative classes are generally lower than in non-quantitative courses.
- For mid-tenure-cycle faculty….

Our review of research related to effective use of student ratings leads us to recommend that:

- Faculty be encouraged to administer online (as well as paper) questionnaires during protected time in class, as this will help maintain high response rates. Online questionnaires should remain open until the final day of classes in order to allow students who were not present in class to complete the questionnaire.
- Faculty who do not administer online questionnaires during class should be encouraged to monitor response rates in real time in order to offer reminders and encouragement in order to produce a high response rate.
- Departments should consider the following issues in their department policies:
    - Whether faculty will be allowed to choose items without departmental approval (we consider this to be a reasonable allowance, as the items behave similarly).
    - By what date faculty need to make the selection of paper administration, since the DAA must print questionnaires at the Print Shop.
    - How to consider class scores when the class size is very small. We recommend that scores always be reported (in RTP binders) as the mean plus/minus one standard deviation, and that a score only be considered below target if the entire range is below the target score. This will allow some information to be gleaned from small samples, but it recognizes the uncertainty of scores derived from small samples.

  - How to consider scores when the response rate is very low. Research on student ratings suggests that very small classes must have a high response rate in order to be reliable, but low response rates don’t have a big impact on the reliability of most student ratings. A commonly used guideline (Nulty, 2008) is presented here:

Class Size	Required Response Rate
10	75%
20	58%
30	48%
40	40%
50	35%
60	31%
70	28%
80	25%
90	23%
100	21%

- Instructors should discuss their student rating reports with their chair and/or mentor as part of the process of improving instruction.
- While we expect the Fresno State SRI scores to be generally similar to IDEA scores, they are not directly comparable as they are different instruments based on two completely different models of how to measure teaching effectiveness. Therefore, as we transition from one system to the other, direct comparisons should be avoided. For those who are mid-tenure cycle, for whom progress is being tracked, we recommend that peer evaluations be used instead of numerical scores on the FSSRI instrument.

In terms of the future administration of student ratings at Fresno State, we recommend that:

- A permanent committee of the Senate be established to oversee student ratings, one with faculty representation by all eight colleges. We suggest that it could be a subcommittee of the Undergraduate Curriculum Committee. The chair of that committee should have expertise in research on assessment of pedagogy, and/or scale construction, and/or statistics. The committee should strive to have all three areas of expertise represented.
- That committee will regularly test revisions to the existing instrument. Revisions can be suggested by any interested member of the Fresno State academic community. The committee will decide whether or not the suggestion merits testing. Testing will consist of the new items being offered in conjunction with existing items for one review cycle so that the new item can be compared in terms of reliability and validity. The committee will make recommendations based on these tests, and changes must be approved by the Senate and the Provost.

- The committee will regularly conduct descriptive analyses of student ratings data to explore areas where Fresno State faculty that could benefit from interventions provided by the Center for Faculty Excellence.

References

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. Psychological Science, 26, 8, 1295-1303. doi:10.1177/0956797615588195

Beran, T. & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? Assessment and Evaluation in Higher Education; 30:6, 593-601.

Cortina, J.M. (1993). What is a coefficient alpha? An examination of theory and application. *Journal of applied psychology, 78, 1,* 98.

Kardas, M. & O'Brien, E. (2018). Easier seen than done: Merely watching others perform can foster an illusion of skill acquisition. Psychological Science, 29, 4, 521-536. doi:10.1177/0956797617740646

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77, 6,* 1121-1134.

MacNell, L, Driscoll, A. & Hunt, A.N. (2014). What in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40, 4,* 291-303.

Nulty, D.D. (2008). The adequacy of response rates to online and paper surveys: What can be done?" *Assessment & Evaulation in Higher Education, 33, 3*: 301-314.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica, 47, 2,* 143-148. doi: 10.1016/0001-6918(81)90005-6.

Uttl, B. & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. Peer J 5:e3299 https://doi.org/10.7717/peerj.3299

Walter, E.M., Henderson, C.R., Beach, A.L., & Williams, C.T. (2016). Introducing the Postsecondary Instructional Practices Survey (PIPS): A concise, interdisciplinary, and easy-to-score survey. CBE Life Sci Educ, December 1, 2016, 15:ar53 DOI:10.1187/cbe.15-09-0193