Academic Affairs Committee
Preliminary Report, Charge AA1:

Re-evaluate operational recommendations 1 through 3 in the "Steps to Establishing an Effective System of Student Ratings" report from 2013, in light of the research data and information collected since the implementation of SmartEvals in 2013.

Subcommittee Members:

Joseph Lanzafame jmlsch@rit.edu
David Halbstein dlhfaa@rit.edu

Background:

In 2013, the Academic Senate endorsed the following four operational recommendations with regard to implementing the SmartEvals system:

1. Use of the SmartEvals system to gather student ratings of teaching effectiveness in classroom settings across the university.
2. Use of the same set of established core items across the university that were used in the pilot ($\alpha = 0.93$ from pilot).
3. Provide the online results for individual instructor (except for instructor-added items) only to the instructor, instructor's immediate supervisor and dean, the provost, and tenure and promotion committees per college guidelines.
4. Re-evaluate recommendations 1-3 after three years of data collection with SmartEvals.

It seems clear that the original charge stems from the fourth "operational recommendation" and seeks new recommendations on the continued use of the system (operational recommendation 1), the questions used within the system (operational recommendation 2), and the dissemination of results (operational recommendation 3).

Such recommendations cannot be made in a vacuum. At the same time as these operational recommendations were made, another recommendation was made[2]. A recommended research plan, included in a supplemental report from the Academic Affairs Committee in March 2013 entitled "Steps to Establishing an Effective System of Student Ratings" report, asked the Provost to designate an individual or group of individuals to research certain aspects of the use of the SmartEvals implementation and report on the results after a period of three years. Suggested components of such a research report would include the following:

1. Monitor for drifts in average ratings attributable to implementation of the new system compared to previous systems.
2. Monitor return rates and association with strategies to improve return rates.
3. Examine effects on ratings of variables associated with course, respondent, instructor, and survey characteristics.

4. Track attitudes, perceptions, and practices regarding the purposes, uses, and value of student input over time among students, faculty, and administrators.
    a. Track student opinion about the value of their input.
    b. Monitor faculty sentiment regarding benefit of student feedback.
    c. Monitor number of faculty supervisors who consult multiple types of evidence in evaluating teaching effectiveness.
5. Observe documented changes in (and perceptions of) instructional effectiveness as associated with the availability and use of professional development and application of student feedback.
6. Apply research findings in formulating recommendations for system modification following a 3-year period of data gathering.

Update on Progress

The Provost's Office interpreted the supplemental report as guidance on what might be collected but not as a mandate that all the components would be addressed and as a result, no formal research plan encompassing all of the 6 components mentioned above was implemented during the interim three years. As such, efforts to flesh out new recommendations have been somewhat hampered by lack of data.

During 2016-2017, the Academic Affairs sub-committee (Joe Lanzafame and David Halbstein) attempted to acquire as much data as we could in an attempt to answer some of the research questions in order to help formulate new recommendations. We worked with the Provost's Office to collect these data.

Research Question 1:

"Monitor for drifts in average ratings attributable to implementation of the new system compared to previous systems."

This simply cannot be done in any meaningful way. Data exists for the current system using the current questions. However, there was no single system being employed before SmartEvals and the questions themselves were different. It is our considered opinion that no meaningful comparison could be made between SmartEvals and the variety of systems it replaced.

Research Question 2:

"Monitor return rates and association with strategies to improve return rates."

The return rates were monitored: 64% in 2013-14, 62% in 2014-15, and 66% in 2015-16. The only significant change during that time frame was the addition of the Student Government questions starting in 2016. It is unclear if this has had an effect on the response rate. The response rate should continue to be monitored.

Research Question 3:

"Examine effects on ratings of variables associated with course, respondent, instructor, and survey characteristics."

A request for data was made to Fernando Naveda, currently the Director of the Office of Intersession and Summer, who oversees SmartEvals. We supplied a number of specific questions regarding course characteristics such as class level and class size, as well as questions regarding instructor characteristics and whether these factors might affect student evaluations. We also asked for data on variability of instructor evaluations over time or across different courses.

As the data analysis was undertaken by Dr. Naveda and Dr. Michael Long, the statistician, it became clear that there were issues associated with the way data was being reported within the SmartEvals system. Some data was obtained. However, one of the main issues that arose from the preliminary data analysis relates to utilization of mean scores and whether reporting mean (average) values for Likert data is appropriate. This is a controversial subject in the literature[4-9] and one that bears further scrutiny. Additionally, the reporting of three significant digits (4.12, e.g.) in the averages clearly overstates the precision with which the responses are known, especially when response rates are only approximately 60%.

There are a number of questions that still need to be addressed:

1. *What statistical measure is appropriate to Likert data?*

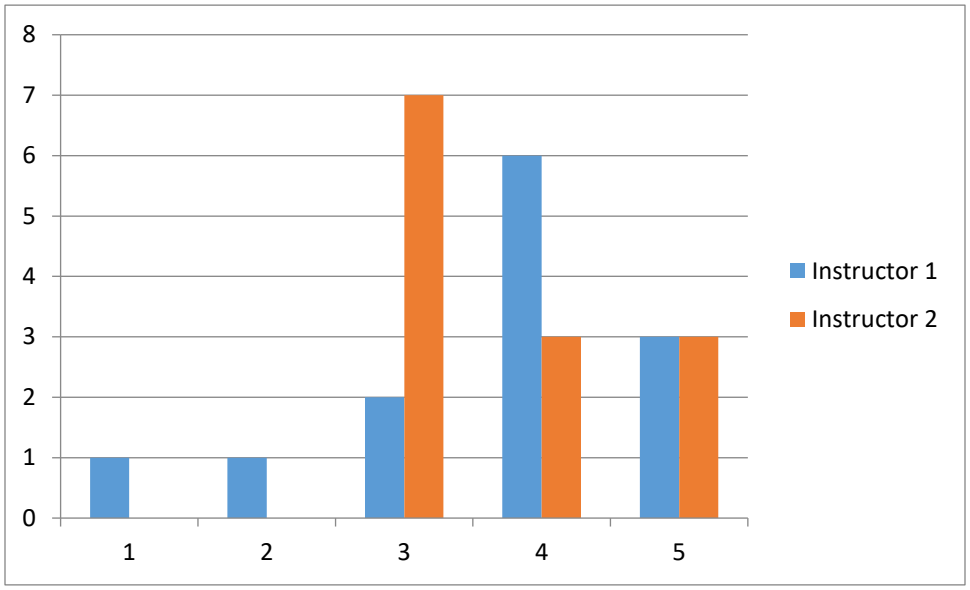
Consider the following statistical examples involving two fictional instructors each teaching a class of 20 students:

On the question "the instructor was an effective teacher", the two instructors received the following responses from 13 of their 20 students which is approximately the average RIT response rate.

Instructor #1		
Likert Score	**Number of Responses**	**Response Totals**
1	1	1
2	1	2
3	2	6
4	6	24
5	3	15
	Avg Score	**3.69**

Instructor #2		
Likert Score	**Number of Responses**	**Response Totals**
1	0	0
2	0	0
3	7	21
4	3	12
5	3	15
	Avg Score	**3.69**

In these two examples, the average score is identical, but there is a clear difference in distribution which can be easily seen in a corresponding bar graph:



The *mean* of both instructors is 3.69. The *median* (preferred for Likert-type data) and mode, however, for Instructor 1 is 4 while it is only 3 for Instructor 2. It would seem that Instructor 1 has, overall, higher evaluations but that is lost in looking at the mean because of a pair of lower that are skewing the data.

2. *What is the actual standard of performance and how do we measure it?*

It is hard to even know what this result means with respect to the actual performance of the two instructors. Both of them have average scores that are BELOW the RIT average, but would we really conclude that either one of them has a performance issue? Instructor 1 has superior evaluations (4 or 5) from 9 of the 13 students. Instructor 2 has fewer superior responses (6 out of 13) but no one rated her lower than 3.

Looking at the median, we now do see a difference between the two instructors with Instructor 1 having a 4, equal to the overall RIT Institute value, while Instructor 2 now has a 3 which is below the RIT Institute as a whole. But how should we interpret the difference? Instructor 2 does seem to have, overall, responses below those of Instructor 1 but those results are not necessarily sub-standard.

3. Is the data that is produced by the SRATE system interpreted correctly?

The numbers produced by the Likert surveys, whether produced by determining the median or the mean, are not meaningful unless evaluated in context. The SRATE system allows for either

or both to be produced, and offers a context for evaluation in the report titled “Percentile Ranking”.

Taken as a raw number, a “4” on a scale of 1 – 5 can (and is) interpreted as a reasonably high score. However, when placed in the context of a percentile ranking among peers, among like courses, or against the institute as a whole, the raw score can take on a different meaning.

In the example below, this instructor has a score of “4” on Question 11, regarding “Clear Communication”. The Percentile Ranking indicates that for this question within the comparison group, is a relatively low score – with nearly 70% of others rated scoring higher.

But, even if this instructor ranks in the bottom 30% on this question, that may not indicate substandard performance. The bottom 30% of passing scores in a class are still passing scores.

There were: 14 possible respondents.
Your average score compared to 1302 student responses

	Question Text	N	Above Average 100% - 70%			Average 69% - 30%				Below Average 29% - 1%		
1	◊ Regularly attended class	10	|-1	------	------	------	------	-----|				
3	◊ CIAS - Commitment to take course	11	|-----	------	------	4.5	------	------	-----|			
5	◊ Course was well organized	11		|-----	------	------	------	------	3.9	-----|		
6	◊ Advanced student understanding	11		|-----	------	------	------	4.3	-----|			
7	◊ Amount of work in course	11		|-----	------	------	3.5	------	-----|			
8	◊ Would recommend course	11	|-----	------	4.5-|							
9	◊ Enhanced interest	11	|-----	------	------	4.4	------	-----|				
10	◊ Material presented in organized manner	11			|-----	------	------	------	------	3.8	-----	|
11	◊ Clear communication	11		|-----	------	------	------	------	4	------	-----	|
12	◊ Positive learning environment	11		|-----	------	------	------	4.4	-----|			
13	◊ Helpful feedback provided	11	|-----	------	------	4.6	------	-----|				
14	◊ Supported student progress	11	|-----	------	------	------	4.5	------	------	-----|		
15	◊ Efective teacher	11	|-----	------	------	------	4.4	------	------	-----|		
18	◊ Instructor was available	11	|-----	------	4.6	------	-----|					
19	◊ Feedback was timely	11		|-----	------	------	------	------	4.1	------	-----	|

The lines extending to either side of the raw numbers represent the “Confidence Interval” of the response mean. The Confidence Interval refers to the repeatability of the score within a broader pool of answers; the wider the range, the more likely that a slight change in the distribution of the answers would have a major impact on the result.

For this professor, in nearly every question, the confidence is quite low that the score would be repeatable with a larger distribution; and the range of results based on slight differences could place him at the very top or very close to the bottom of his percentile rank.

To use raw numbers, whether arrived at via median OR mean, without the context of the percentile ranking and the confidence interval associated with that rank would be inaccurate and, if used in determining rank or salary, grossly unfair.

SRATE data is being used for evaluation of faculty effectiveness, yet a full understanding of what a full SRATE report represents may not be fully understood by those making the

evaluations. In our anecdotal experience, the full data set of report numbers has not been requested as part of the normal evaluative process.

4. Do differences in semantic interpretations influence student responses?

The average student response for any particular instructor and any particular question generally has a large standard deviation (approximately 1.0). Anecdotal evidence suggests a significant likelihood that at least some of this deviation is due to the somewhat vague nature of the question and the lack of guidance as to what the responses are supposed to indicate. For example, it is quite likely that two different students in the same class would interpret "The instructor is organized" to mean two different things. And, even if they agreed on what "organized" meant, they might view a "3" to mean something different; to one student a 3 out of 5 would be a failing grade ("60%"), to another a 3 out of 5 would be an average grade.

Research Question 4:

"Track attitudes, perceptions, and practices regarding the purposes, uses, and value of student input over time among students, faculty, and administrators."

a. Track student opinion about the value of their input.

b. Monitor faculty sentiment regarding benefit of student feedback.

c. Monitor number of faculty supervisors who consult multiple types of evidence in evaluating teaching effectiveness.

Student opinion about SRATE was measured in the 2016 administration of the Noel Levitz Student Satisfaction Survey. This survey is administered to all RIT students. In response to the question: "Students have sufficient opportunity to evaluate faculty teaching through the online system "SmartEvals", students indicated an importance score of 5.77 and a satisfaction score of 5.60 on a .7.0 scale. A smaller the gap between the two scores, the greater the confidence level that student satisfaction is at an appropriate level. No progress has been made in addressing faculty sentiment or faculty supervisor practices. A survey of all stakeholders is probably the most reasonable approach to gathering such information. The survey has been delayed due to the recommendations and actions discussed below.

Research Question 5:

"Observe documented changes in (and perceptions of) instructional effectiveness as associated with the availability and use of professional development and application of student feedback."

This question has not been addressed and it is rather challenging to undertake this study in a robust manner. If the sole evaluation of "instructional effectiveness" is taken to be SmartEvals, then it is simple to observe and document changes in student survey responses. However, correlating any such changes to specific professional development and application of student feedback would be more difficult to accomplish. Further, "instructional effectiveness" should not, according to current RIT policy, be judged solely by student survey responses which makes

the determination and evolution of effectiveness more challenging to ascertain. Further, studies exist which suggest a negative correlation is possible between survey responses and actual effectiveness[10,11].

Recommendations

In light of the preliminary results obtained above and conversations with the Provost's Office we make one main recommendation:

Create a Research Committee to further collect and analyze data related to the Research Questions above and to formulate recommendations for improvement to the system and the use of the system.

It should be pointed out that the Provost is already supportive of such an effort. Further, in light of the discussions already undertaken, the Provost has already reminded the Dean's Council that faculty evaluations are not to be solely based on SmartEvals and asked the Deans to inform their administrators that they should not base annual evaluations solely on SmartEvals.

We make the following specific recommendations moving forward:

1. The Academic Affairs Committee of the Academic Senate should maintain a central role, in concert with the Provost's Office, in the creation and implementation of the Research Committee.
2. The Research Committee should investigate appropriate statistical metrics for interpretation of SmartEvals results.
3. The Research Committee should undergo thorough training on the SRATE system as an integral part of their charge, and make recommendations as to the requirement, depth and frequency of training for deans, department chairs, and faculty.
4. The Research Committee should determine objective standards for acceptable performance that do not rely simply on being above or below the Institute average.
5. The Research Committee should survey the faculty to gauge their attitude and responsiveness to SmartEvals along with their understanding of the meaning of the results.
6. Student Government should be involved in the formulation of new guidelines.
7. Consideration should be given to mid-semester formative student evaluations. One major shortcoming of the current system is that any instructional issues that arise are not evaluated until after the semester has concluded. As a result, students will inevitably feel that the faculty are not responsive to their concerns because they do not observe any of the changes that their feedback helps to bring about.
8. The Research Committee should investigate a general system of instructional evaluation that does not rely solely on SmartEvals but that provides administrators with an efficient means of evaluating faculty performance and aiding faculty professional development.
9. While new standards are being researched and implemented, we suggest that SmartEvals data can be used robustly for formative assessment but cautiously for summative assessment.

References:

1. https://digitalarchive.rit.edu/xmlui/bitstream/handle/1850/18759/ASMinutesAugust252016Approved_09-08-2016.pdf?sequence=1
2. https://www.rit.edu/academicaffairs/sites/rit.edu.academicaffairs/files/docs/supplementalreportestablishingeffectivesystemonstudentratings_03-21-2013.pdf
3. Emery, Charles R., Kramer, Tracy R., Tian, Robert G. "Return to academic standards: a critique of student evaluations of teaching effectiveness." *Quality Assurance in Education* 11:1 (2003): 37-46
4. Stark, Philip B., and Richard Freishtat. "An evaluation of course evaluations." *ScienceOpen Research* 9 (2014): 2014.
5. Marsh, Herbert W. "Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness." *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Springer Netherlands, 2007. 319-383.
6. Lubke, Gitta H., and Bengt O. Muthén. "Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons." *Structural Equation Modeling* 11.4 (2004): 514-534.
7. Jamieson, Susan. "Likert scales: how to (ab) use them." *Medical education* 38.12 (2004): 1217-1218.
8. Allen, I. Elaine, and Christopher A. Seaman. "Likert scales and data analyses." *Quality progress* 40.7 (2007): 64.
9. Norman, Geoff. "Likert scales, levels of measurement and the "laws" of statistics." *Advances in health sciences education* 15.5 (2010): 625-632.
10. Braga, Michela, Marco Paccagnella, and Michele Pellizzari. "Evaluating students' evaluations of professors." *Economics of Education Review* 41 (2014): 71-88.
11. Carrell, Scott E., and James E. West. "Does professor quality matter? Evidence from random assignment of students to professors." *Journal of Political Economy* 118.3 (2010): 409-432.