Academic Affairs Committee

Charge 1 – SmartEvals

Background: Charge AA1

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.

Research questions

- We formulated a number of research questions based on recommendations that were made at the same time as the charge.
- We began by using those questions as a foundation for the re-evaluation of SmartEvals.
- We undertook to gather data around those research questions…and then the earth shook.

Research questions

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”

This question proved to be the most salient, and the most troublesome

Are evaluation results affected by:

Class size?	YES
Class level?	YES
Class format?	YES
Student grade?	YES

Is there consistency in evaluations for an individual instructor:

From one class to the next?	NO
From one semester to the next?	NO

The issue with the data…

The inconsistency of the results was consistent.

We found that the *nature of the data* itself presented the major obstacle.

Appropriate use of Likert data:

Mean or Median?

Can the Likert Scale data be condensed into a single, meaningful "grade"?

Is there a difference when using the mean vs median? Which is correct?

How many significant figures are meaningful?

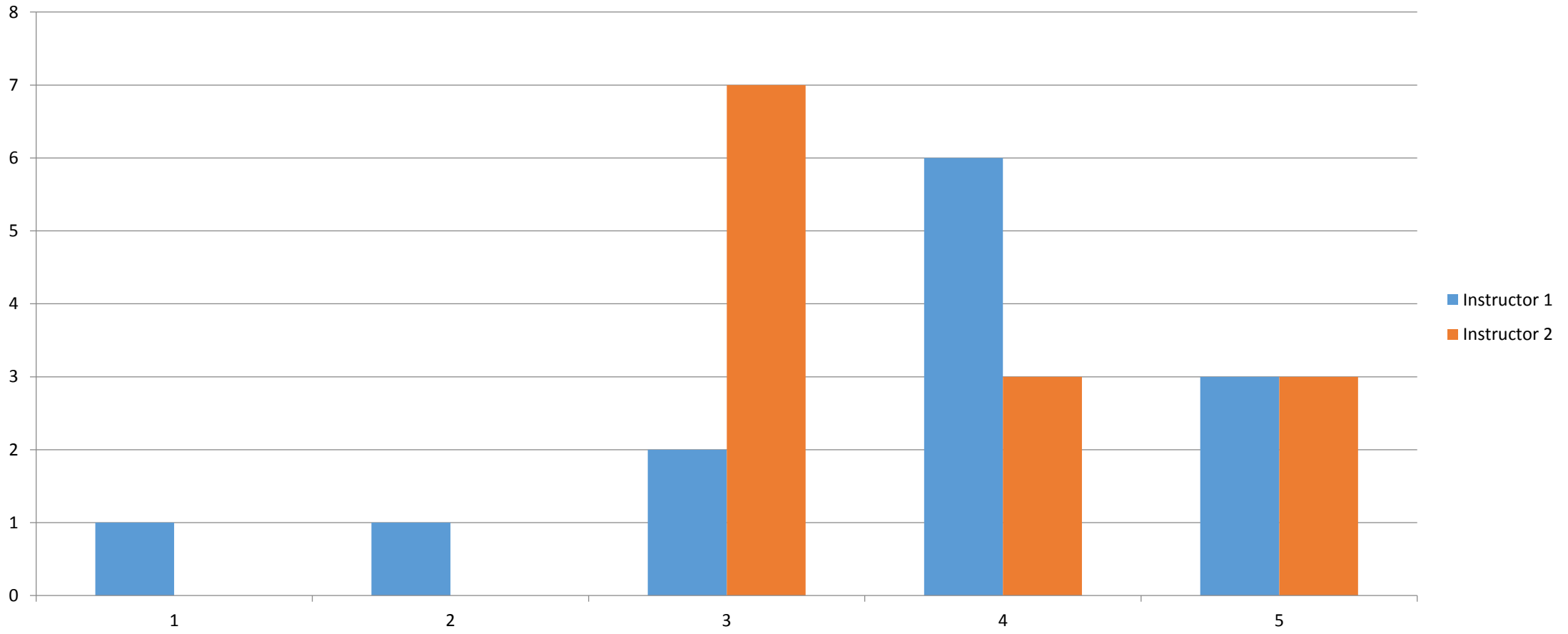
Is there really a difference between a 3.87 and a 4.21?

Are we *observing* distinctions or *manufacturing* distinctions?

2 “identical instructors” – 13 opinions in a class of 20 students

	Instructor #1				Instructor #2	
Likert Score	Number of Responses	Response Totals		Likert Score	Number of Responses	Response Totals
1	1	1		1	0	0
2	1	2		2	0	0
3	2	6		3	7	21
4	6	24		4	3	12
5	3	15		5	3	15
	Avg Score	3.69			Avg Score	3.69

2 “identical” instructors?



What is the standard?

Given the limitations in the data itself, it is unclear what mere comparison to the “average” indicates.

- *What is the objective standard?*
- *How do we assess whether individual faculty are meeting that standard?*

There are no clear answers to these questions.

What the chart means…

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	------	-----|	

With a confidence interval of 95%, this range shows the likelihood of repeated performance data within a larger sample group.

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	------	-----	|

In many cases (as pictured above), the predicted scores can vary across the entire range as sample size increases.

In addition to the challenges related to the data after it has been collected, there are certain specific challenges related to the collection itself.

The questions themselves are subject to individual semantic interpretation, which contributes to another level of variability:

19: What does “timely” mean?

#5: What does “organized” mean?

18: What is the definition of “available”

For that matter,

What does “3” mean? Is mine the same as yours?

OUR RECOMMENDATION

Create a Research Committee:

- ***to further collect and analyze data related to the Research Questions above***
- ***to formulate recommendations for improvement to the system and to formulate recommendations as to the appropriate use of the system.***

The Provost has already agreed to undertake this effort.

Our Recommendations:

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.

Our Recommendations (continued):

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.

Thank You!

A huge THANK YOU to…

1. The entire AAC – what a great group!
2. Penchapakesan Venkataraman – AAC chair
3. Christine Licata and Provost Jeremy Haefner for continued constructive dialogue.
4. Fernando Naveda and Mike Long

Questions?

Research questions and results:

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. The data does not exist.

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.

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

Data was collected but the statistical issues cloud the interpretation.

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

1. Track student opinion about the value of their input.
2. Monitor faculty sentiment regarding benefit of student feedback.
3. 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. We did not endeavor to survey the faculty and administrators once it had been decided to reevaluate the entire system.

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.

There is no clear mechanism by which to do this.

6. Apply research findings in formulating recommendations for system modification following a 3-year period of data gathering.