Analyses of the Reliability of Student Ratings of Teaching at Kennesaw State University using Digital Measures Course Response: Spring 2012 through Spring 2014

Report prepared by Thomas Pusateri, CETL Associate Director for SoTL
Updated October 13, 2014

KSU began using Digital Measures Course Response in Fall 2010 to collect student ratings online for all courses. As of Spring 2014, KSU distributed a total of 762,350 electronic forms to students in 30,854 classes, saving paper and staff processing time compared to the previous paper forms.

In Spring 2012, the KSU Faculty Senate approved the inclusion of two multiple-choice items on all ratings forms, one item on *course content* and one on *instructor effectiveness*. Students rated each of these items on a 4-point scale (1=Strongly disagree, 2=Disagree, 3=Agree, 4=Strongly agree):

- "Overall the content of this course contributed to my knowledge and intellectual skills."
- "The instructor was effective in helping me learn."

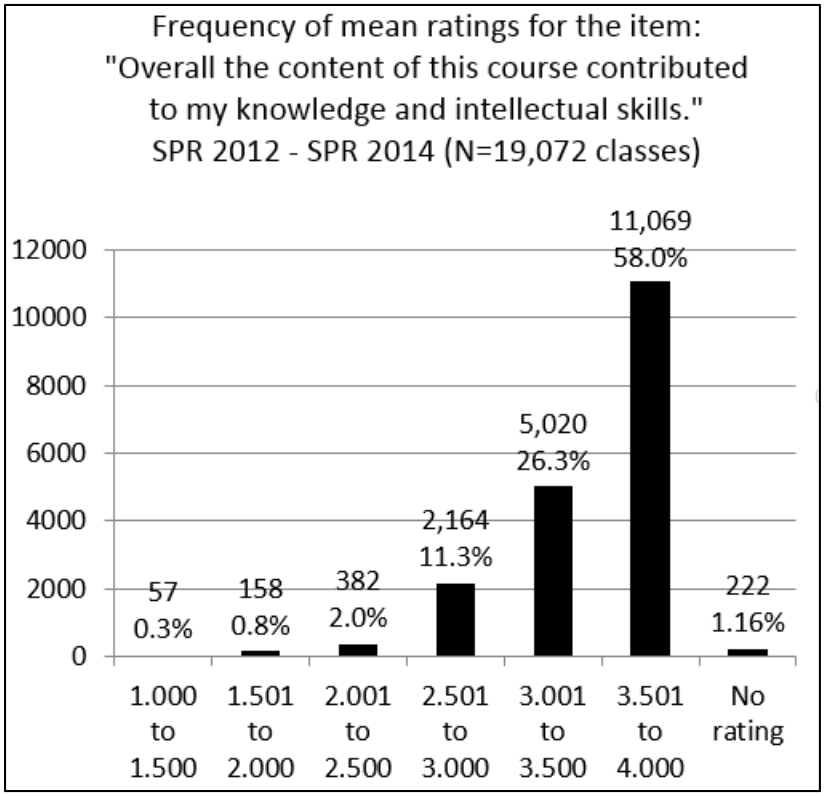
This report summarizes analyses of the reliability of student feedback on these two items collected from Spring 2012 through Spring 2014. During these semesters, KSU distributed a total of 481,013 electronic forms to students in 19,072 classes. Students visited the Digital Measures Web site to provide responses to a total of 171,023 of these forms, for an overall response rate of 35.6%. Students who visited the Web site also had the option to decline completing a form. During these semesters, students actively declined to complete 6,427 forms; as a result, the total number of completed (answered + declined) forms is 177,450, for an overall completion rate of 36.9%.

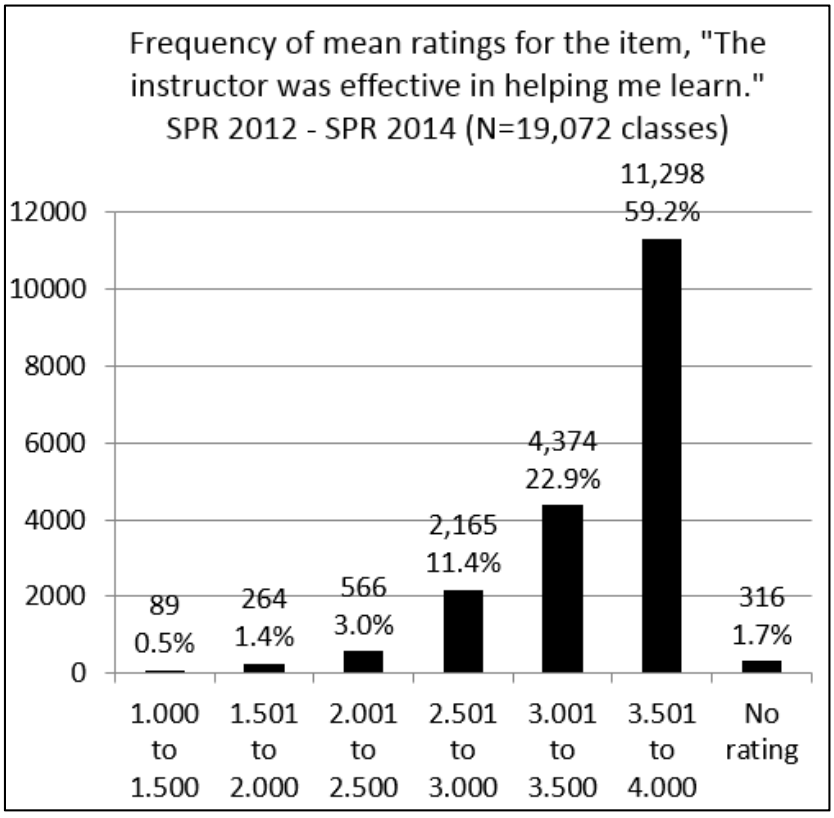
The table below indicates the distribution of student ratings on the two university-wide items for all ratings collected from Spring 2012 through Spring 2014.

DISTRIBUTION OF STUDENT RATINGS ON THE TWO UNIVERSITY-WIDE ITEMS		No Response	1-Strongly Disagree	2-Disagree	3-Agree	4-Strongly Agree	Total
Overall the content of this course contributed to my knowledge and intellectual skills.	# of ratings	1,947	5,600	8,973	49282	102,967	168,769
	Percent	1.15%	3.32%	5.32%	29.20%	61.01%	100.00%
The instructor was effective in helping me learn.	# of ratings	2,985	6,771	10,128	42872	105,833	168,589
	Percent	1.77%	4.02%	6.01%	25.43%	62.78%	100.00%

Students have given high average ratings to both *course content* and *instructor effectiveness*. Approximately 90% of students responded "Agree" or "Strongly Agree" to each of the items.

Across classes, the average (mean) ratings of *course content* and *instructor effectiveness* have also been high. The tables on page 2 depict the distribution of average ratings (using the arithmetic mean) provided by students in each of the 19,072 classes reviewed. Over 60% of classes received average ratings for *course content* and *instructor effectiveness* of 3.5 or higher on the 4-point scale (midway between 3=Agree and 4=Strongly Agree), and over 80% of classes received average ratings of 3.0 or higher (3=Agree).


Frequency of mean ratings for the item:
"Overall the content of this course contributed to my knowledge and intellectual skills."
SPR 2012 - SPR 2014 (N=19,072 classes)
12000
10000
8000
6000
4000
2000
0
57
0.3%
158
0.8%
382
2.0%
2,164
11.3%
5,020
26.3%
11,069
58.0%
222
1.16%
1.000 to 1.500
1.501 to 2.000
2.001 to 2.500
2.501 to 3.000
3.001 to 3.500
3.501 to 4.000
No rating


Frequency of mean ratings for the item, "The instructor was effective in helping me learn."
SPR 2012 - SPR 2014 (N=19,072 classes)
12000
10000
8000
6000
4000
2000
0
89
0.5%
264
1.4%
566
3.0%
2,165
11.4%
4,374
22.9%
11,298
59.2%
316
1.7%
1.000 to 1.500
1.501 to 2.000
2.001 to 2.500
2.501 to 3.000
3.001 to 3.500
3.501 to 4.000
No rating

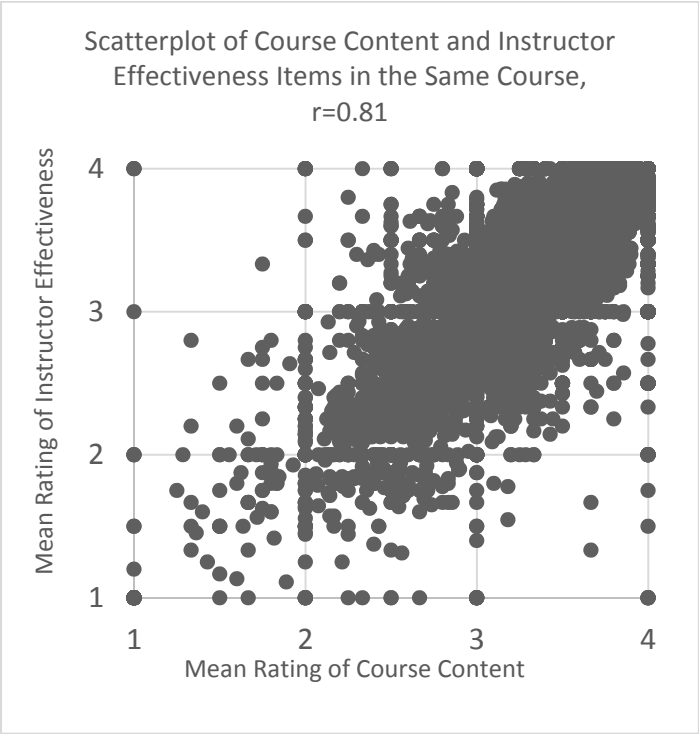
- **RELIABILITIES OF *COURSE CONTENT* AND *INSTRUCTOR EFFECTIVENESS* ITEMS**

The table below provides results from analyses mean student ratings of the *course content* and *instructor effectiveness* items across all classes taught from Spring 2012 through Spring 2014. The analysis in the first row includes all classes in which students responded to both the *course content* and *instructor effectiveness* items. The remaining rows display the results of three separate analyses for *instructor effectiveness* and three for *course content*. The first analysis compares mean ratings in classes in which the same instructor taught two classes of the same course in the same semester via the same modality (e.g., both face-to-face, both hybrid, both online), the second compares classes in which the same instructor taught two different classes in the same semester via the same modality, and the third compares the same course taught by two different instructors in the same semester via the same modality.

Analyses in the left columns include all classes in which at least one student provided a rating. Analyses in the right columns include classes in which at least five students provided a rating.

RELIABILITY ANALYSES FOR CLASSES TAUGHT IN THE SAME SEMESTER VIA THE SAME MODALITY (Both classes were taught F2F or Hybrid or Online)	Classes with at least 1 student rating		Classes with at least 5 student ratings	
	Number of classes	Correlation	Number of classes	Correlation
Correlation among ratings of Instructor Effectiveness and Course Content within the same class	18729	0.81	12498	0.86
RELIABILITY OF RATINGS OF INSTRUCTOR EFFECTIVENESS	Number of pairs of classes	Correlation	Number of pairs of classes	Correlation
Same instructor, Same course	3580	0.57	2641	0.63
Same instructor, Different course	4358	0.34	2813	0.41
Different instructor, Same course	6084	0.13	4333	0.10
RELIABILITY OF RATINGS OF COURSE CONTENT	Number of pairs of classes	Correlation	Number of pairs of classes	Correlation
Same instructor, Same course	3614	0.48	2664	0.54
Same instructor, Different course	4409	0.28	2858	0.33
Different instructor, Same course	6139	0.19	4364	0.15

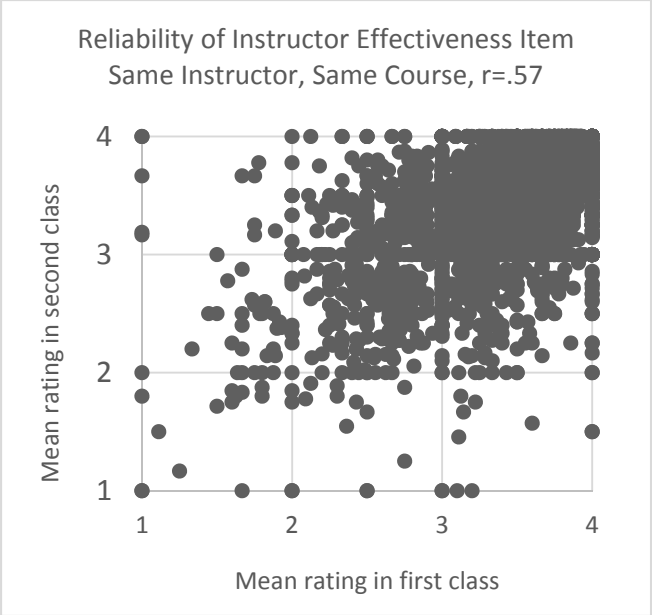
Students appear to be responding similarly to the *course content* and *instructor effectiveness* items. The high correlation (0.81 overall, 0.86 for classes with 5 or more student ratings) between these items indicates that students tended to give similar ratings to both items. This suggests that students were not differentiating between the course content and the instructor who taught the course. Other sections of this report will discuss this further.

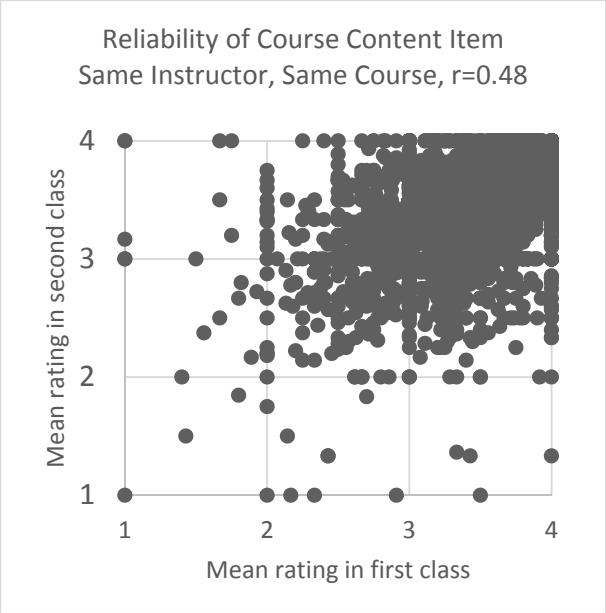


Based on comparisons of ratings provided by students in classes of the same course taught by the same instructor in the same semester via the same modality, the reliability of the *instructor effectiveness* and *course content* items are lower than desirable[1]**.** The correlation of mean ratings on the *instructor effectiveness* item is 0.57 (and increases to 0.63 when for classes where at least 5 students responded). The correlation of mean ratings of the *course content* item is lower (0.48 for all classes, 0.54 for classes in which at least 5 students responded). These correlations are lower than similar correlations reported in the literature.[2]

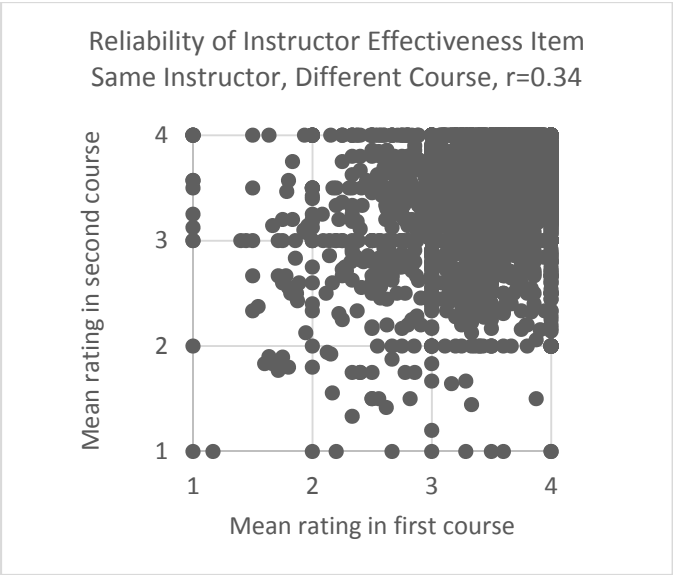
[1] In Chapter 4 of the *Handbook of Statistics, Vol. 26* (2006), Webb, Shavelson, and Haertel indicate that: "[Reliability c]oefficients at or above 0.80 are often considered sufficiently reliable to make decisions about individuals based on their observed scores, although a higher value, perhaps 0.90, is preferred if decisions have significant consequences. Of course, reliability is never the sole consideration in decisions about the appropriateness of test uses or interpretations."

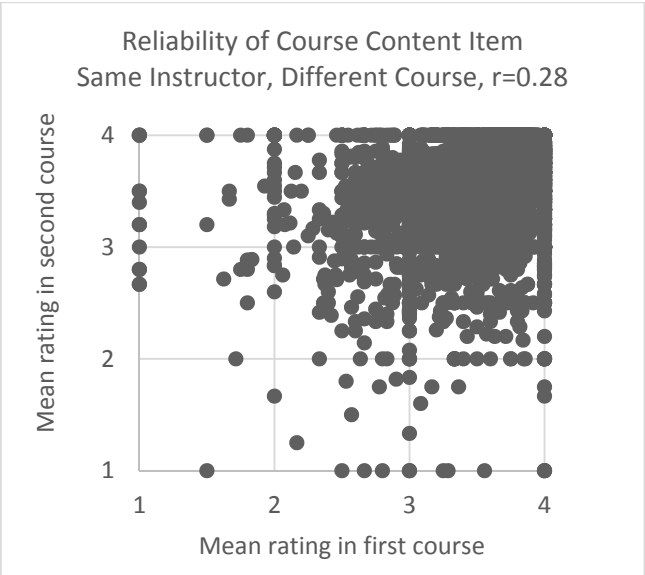
[2] For example, Marsh (1984) reported a correlation of 0.71 on instructor items and 0.69 on content items when comparing student ratings in the same course taught by the same instructor in the same semester. Marsh's data involved ratings forms that included multiple items for rating both the instructor and course content, whereas KSU's form includes one instructor item and one course content item. Including more items on a form typically increases the correlations obtained in analyses of reliability.
Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.





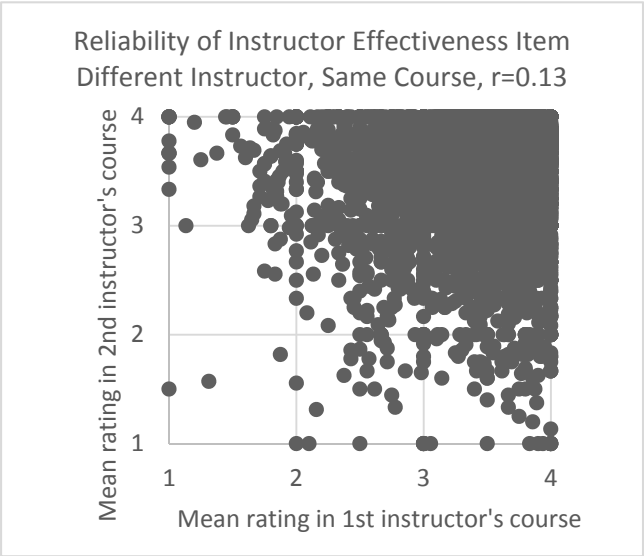
The reliability of both items declines when comparing the ratings provided by students in different courses taught by the same instructor in the same semester via the same modality. The correlation of the *instructor effectiveness* item drops from 0.57 (same course) to 0.34 (different courses), and the correlation of the *course content* item drops from 0.48 to 0.28. This result is not unexpected.[3]

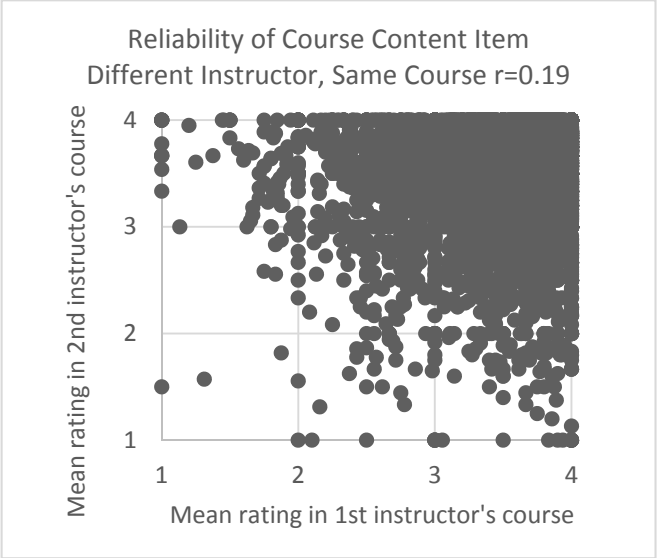




[3] Marsh (1984) reported a drop from 0.71 to 0.52 on instructor items and from 0.69 to 0.34 on content items when comparing student ratings in the same course versus a different course taught by the same instructor. Marsh's data involved ratings forms that included multiple items for rating both the instructor and course content, whereas KSU's form includes one instructor item and one course content item.
Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707-754.

The reliability of both items declines further (and is nearly zero) when comparing the ratings provided by students taking the same course in the same semester via the same modality from different instructors. The correlation of the *instructor effectiveness* item drops to 0.13 (0.10 for classes in which at least 5 students responded). The correlation of the *course content* item drops to 0.19 (0.15 for classes in which at least 5 students responded).





The *instructor effectiveness* item is not behaving as well as we should expect. Students do provide more similar ratings for the same instructor in different sections of the same course taught via the same modality, and they provide less similar ratings for different instructors teaching the same course. However, the correlation between mean ratings of the same instructor teaching the same course in the same semester is lower that similar correlations reported in the research literature (refer to Footnotes 2 and 3).

The *course content* item is also not behaving as well as we should expect. Ideally, students should be responding to the content of the course, not the instructor, which should result in a higher correlation than obtained when the same course is taught by different instructors. Although some courses may be standardized, instructors may have sufficient flexibility to teach the same course in ways that are sufficiently different from other instructors that students may be responding to the *course content* item based on the instructor's approach in that class.

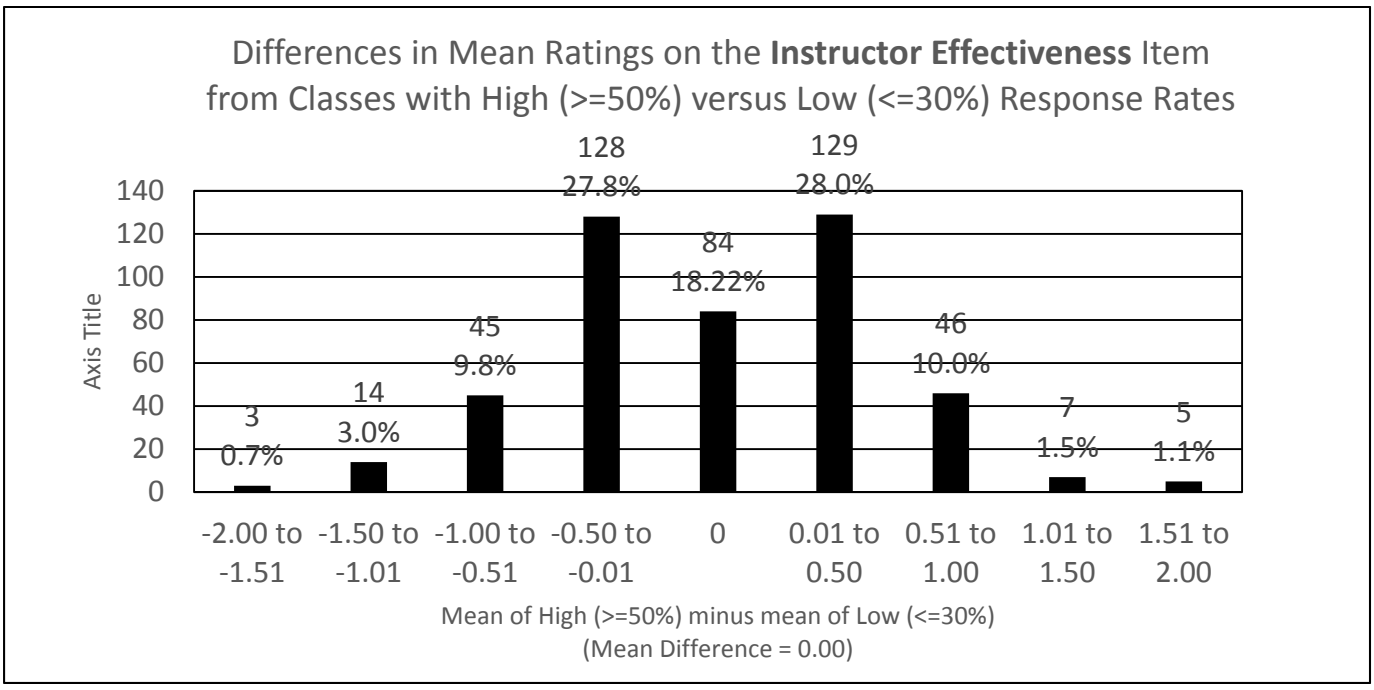
• RELIABILITY AND HIGH VERSUS LOW RESPONSE RATES

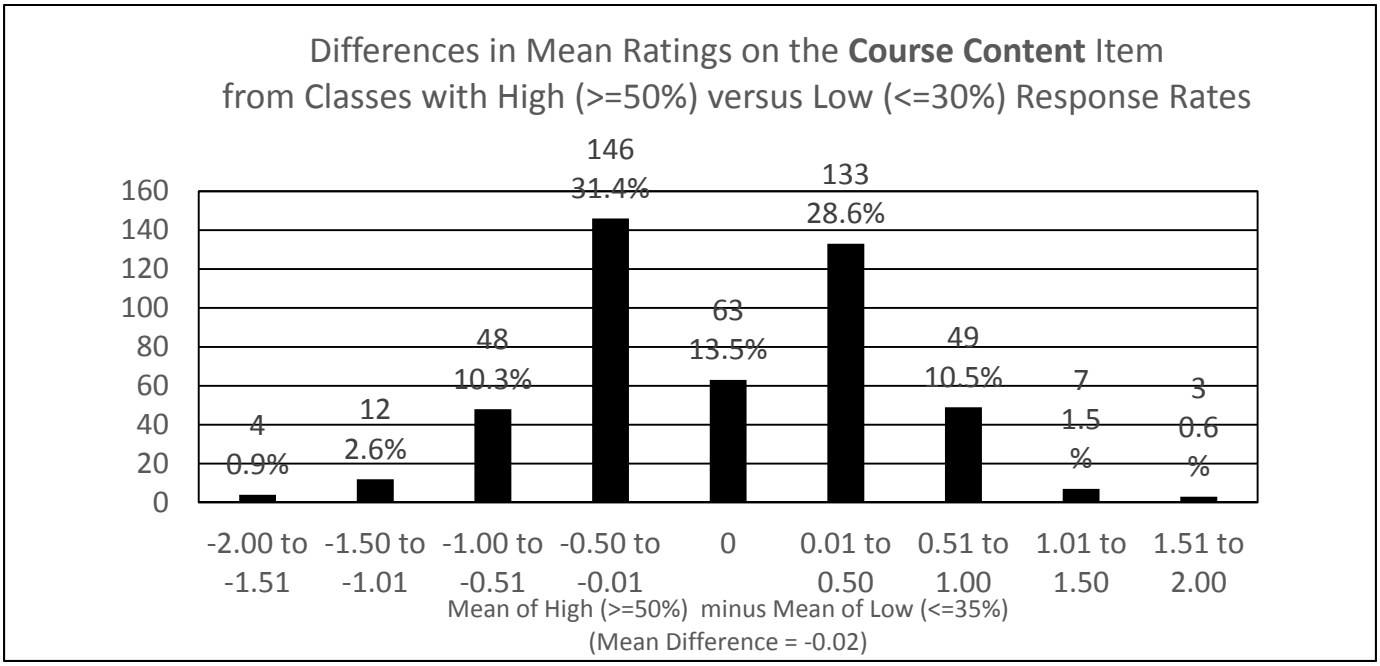
The response rate of 35.6% is sufficiently low that it may raise concerns for faculty that the feedback they are receiving from students in classes with low response rates does not sufficiently represent the whole class. The analyses reported below address this question. These analyses compare the ratings provided by students in pairs of classes of the same course taught by same instructor in the same semester via the same modality but that differ in response rates. For each pair of classes chosen, the response rate in one class was 35% or lower and the response rate in the other class was 50% or higher. The analyses in the left columns include all pairs of classes that meet these criteria, and the analyses in the right column restrict the analyses to pairs in which at least 5 students provided ratings in each class.

RELIABILITY OF RATINGS BETWEEN CLASSES TAUGHT BY THE SAME INSTRUCTOR WITH DIFFERENT RESPONSE RATES IN EACH CLASS	Correlations of mean class ratings between pairs of classes of the same course taught during the same semester by the same instructor, one class with a response rate of 50% or higher and the other class with a response rate of 35% or lower.			
	Classes with at least 1 student rating		Classes with least 5 student ratings	
RELIABILITY OF RATINGS OF INSTRUCTOR EFFECTIVENESS	Number of pairs of classes	Correlation	Number of pairs of classes	Correlation
Same semester/instructor/course/modality 50% or more responses in 1 class 35% or fewer responses in 1 class	461	0.47	273	0.50
RELIABILITY OF RATINGS OF COURSE CONTENT	Number of pairs of classes	Correl.	Number of pairs of classes	Correl.
Same semester/instructor/course/modality 50% or more responses in 1 class 35% or fewer responses in 1 class	465	0.39	274	0.45

The reliability of both the *instructor effectiveness* and *course content* items decline when comparing sections of courses with low (<=35%) and high (>=50%) response rates. On the *instructor effectiveness item*, the correlation drops from 0.57 (when all response rates are included in the analysis) to 0.47 (when classes with low and high response rates are compared). Similarly, the correlation for the *course content* item drops respectively from 0.48 to 0.39.

Despite the lower correlations, the mean ratings in classes with low (<=35%) and high (>=50%) response rates were similar for both the *course content* and *instructor effectiveness* Items. The charts below display the distribution of differences in the ratings on each item, subtracting the mean rating from classes with low response rates (<=30%) from mean ratings from classes with high response rates (>=50%). The distributions for each item are nearly symmetrical. Students in courses with low response rates were equally likely as students in courses with high response rates to give higher (or lower) ratings on either item.





• RELIABILITY IN FACE-TO-FACE, HYBRID, & ONLINE CLASSES

The table below provides separate analyses of the reliability of student ratings of the *course content* and *instructor effectiveness* items for **face-to-face, hybrid, and online classes** taught from Spring 2012 through Spring 2014.

RELIABILITY ANALYSES: FACE-TO-FACE, HYBRID, AND ONLINE CLASSES	Face-to-Face		Hybrid		Online	
	Number of classes	Correl.	Number of classes	Correl.	Number of classes	Correl.
Correlation among ratings of Instructor Effectiveness and Course Content in each class.	16688	0.81	752	0.85	1289	0.80
RELIABILITY OF RATINGS OF INSTRUCTOR EFFECTIVENESS	Number of pairs of classes	Correl.	Number of pairs of classes	Correl.	Number of pairs of classes	Correl.
Same instructor, Same course	3214	0.57	122	0.61	244	0.58
Same instructor, Different course	4121	0.34	90	0.17	147	0.47
Different instructor, Same course	5696	0.14	138	-0.09	250	0.13
RELIABILITY OF RATINGS OF COURSE CONTENT	Number of pairs of classes	Correl.	Number of pairs of classes	Correl.	Number of pairs of classes	Correl.
Same instructor, Same course	3248	0.48	122	0.47	244	0.45
Same instructor, Different course	4171	0.27	90	0.30	148	0.37
Different instructor, Same course	5747	0.18	139	0.04	253	0.39

The reliability of the *instructor effectiveness* item appears to be unaffected by the mode in which the course is delivered. The correlations for the same instructor teaching the same course in the same semester using the same delivery mode are 0.57 for face-to-face classes, 0.61 for hybrid classes, and 0.58 for online classes.

The correlation for the *course content* item differs for online classes when compared to the other delivery modes. For online classes, the correlation in ratings of *course content* from students taking the same course in the same semester taught by different instructors is 0.39. This is higher than the correlations obtained in face-to-face (0.18) and hybrid (0.04) classes, and it is similar to the correlation for students taking different courses online from the same instructor (0.37). A possible explanation for these results is that instructors of online courses must first complete training in online course development, which may contribute to greater similarity in course content across courses taught by the same instructor and across course sections taught online by different instructors.

• ADDENDUM: REQUESTED ANALYSES (9/24/2014)

During the September 22, 2014 meeting of the Chairs' and Directors' Assembly, one chair questioned whether the correlations obtained for the *course content* and *instructor effectiveness* items (obtained for courses in which the same instructor taught two different classes in the same semester) might be an artifact of the highly skewed distributions of mean ratings. The chair was concerned that approximately 90% of student ratings were "Agree" or "Strongly Agree" which the chair thought might produce correlations of means by chance. To test this concern, I reran the analyses of that data using the same distribution of means for the *course content* and for the *instructor effectiveness* items but randomly associating each mean with different courses. I then reran the analysis a second and third time, each time using a different random order of the same means. As you can see in the table below, despite the skewed distributions of mean ratings, the correlations of random means is zero. The correlation of the actual means is not an artifact of the skewed distributions of means.

RELIABILITY ANALYSES FOR CLASSES TAUGHT IN THE SAME SEMESTER	Classes with at least 1 student rating		Classes with at least 5 student ratings	
	Number of classes	Correlation	Number of classes	Correlation
Correlation among ratings of Course Content and Instructor Effectiveness within the same class				
ACTUAL MEANS	18729	0.81	12498	0.86
Randomly-ordered means: Trial 1	18541	0.01	12559	0.01
Trial 2	18536	0.00	12568	0.00
Trial 3	18537	0.00	12531	0.00
RELIABILITY OF RATINGS OF INSTRUCTOR EFFECTIVENESS SAME INSTRUCTOR, SAME COURSE	Number of pairs of classes	Correlation	Number of pairs of classes	Correlation
ACTUAL MEANS	3580	0.48	2664	0.54
Randomly-ordered means: Trial 1	3669	0.01	2808	0.01
Trial 2	3663	-0.02	2807	-0.02
Trial 3	3665	-0.02	2795	-0.03
RELIABILITY OF RATINGS OF COURSE CONTENT SAME INSTRUCTOR, SAME COURSE	Number of pairs of classes	Correlation	Number of pairs of classes	Correlation
ACTUAL MEANS	3580	0.57	2641	0.63
Randomly-ordered means: Trial 1	3705	-0.03	2830	-0.02
Trial 2	3719	0.01	2836	0.01
Trial 3	3697	0.03	2830	0.02