campuslabs

DATA IN ACTION SERIES

Myths and Realities: An Examination of Course Evaluations in Higher Education

AUTHORS

Will Miller, Ph.D.
Executive Director, Institutional Analytics, Effectiveness and Strategic Planning
Jacksonville University

Tyler Rinker, Ph.D.
Manager, Data Science
Campus Labs

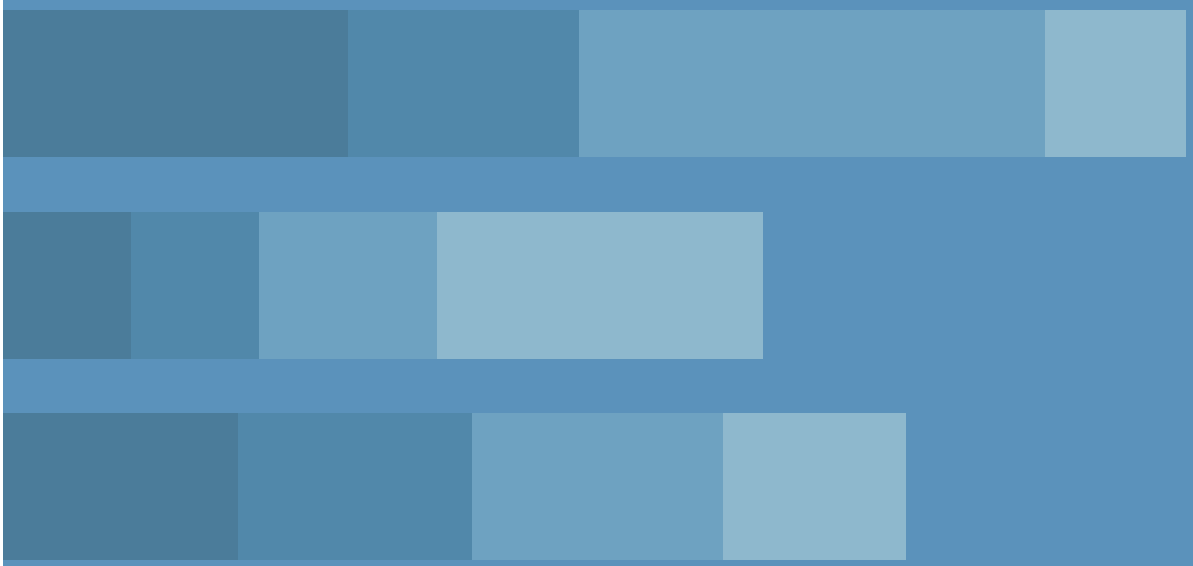


Table of Contents

Introduction 3

Students who take course evaluations outside of class time are more likely to be critical in their comments and ratings 9

Comments on course evaluations only reflect the extremes and consequently are not constructive 12

Course evaluation instruments do not accurately measure what faculty and administrators want 15

Low response rates skew course evaluation results 24

Respondents have a consistent attitude across different evaluations 26

Evaluation comments can be a predictor for average rating on course evaluations 28

Conclusion and suggested areas for further research 31

Campus Labs Data Science 32

About the authors 33

Introduction

One topic of longstanding debate in higher education is that of course evaluations. While the questions, scales, methods for administration and weights applied to student ratings of instruction may vary significantly from campus to campus, the political nature of course evaluations and concerns about student biases are fairly consistent. Taken at face value, the entire concept of course evaluations or faculty effectiveness surveys strikes at another topic of considerable debate in the industry—that of students being viewed as consumers. In many instances, faculty ask what qualifies a student to judge whether a course or faculty member has been pedagogically effective in relaying essential information and skills, while administrators often highlight the role students serve as consumers being asked to rate a delivered service.

While this divide alone is enough to warrant further study—measured biases make the topic even more important for discussion and analysis. These biases—whether focused on race, gender or other factors associated with an instructor—are typically implicit, which makes them problematic to correct for since a student likely is unaware of the underlying causes of their ratings. But for women and minority faculty, there is a legitimate assumption that they are held to a higher standard than their white, male counterparts—another area that necessitates continued examination in our industry. In an effort to overcome these implicit biases, many women and minority faculty often devote greater energy to teaching, adding burden on top of already great research and service commitments—all of which has possible impact on tenure, promotion and salary decisions.

As stated, this ongoing debate on course evaluations is not new. In fact, the preponderance of evidence suggests that campuses have been aware of many of these concerns for quite some time. So, if this is the case, why do course evaluations continue to occur on nearly every campus across the country? It starts with the reality that students are the only individuals who observe faculty teaching on a daily basis throughout a semester. While they may bring biases to their assessments, who else on campus has a reservoir of classroom experiences with a faculty member to levy any type of assessment on instructional effectiveness?

Many faculty have argued they do not object to the concept of course evaluations, but instead take issue with how they oftentimes stand alone as unitary measures of pedagogical excellence as opposed to being part of a more holistic view. Even as many institutions do utilize more holistic mechanisms for evaluating faculty effectiveness, they still include course evaluations, noted potential biases and all, as part of that overall evaluation—what, then, does that say for higher education as a whole? The lack of a viable standalone, holistic alternative complicates any efforts to discontinue course evaluations in the near future. And, when weighted for bias, and evaluated with a human eye, course evaluations can provide great value and an immense amount of usable data.

Institutions could decide to pre- and post-test students to show learning growth. But that would require students to take such assessments seriously, faculty to not teach to the test, and everyone to agree every type of learning is quantifiable and equitable—which is not likely to happen. Faculty regularly advocate for self-reporting yet rarely are willing to meaningfully criticize their own approach and performance in more than a cursory manner. Some have suggested student performance in subsequent courses could be a useful proxy for course evaluations—and in an era of curriculum maps and student learning outcomes, such measures are useful to track. But does the faculty member in an initial course really bear the burden of how a student does later in their academic program? Should they be expected to attach their name in that way?

Likely not. Moreover, in small programs they may be the very same faculty teaching later courses—or, survivor bias may come into play.

It could be possible to evaluate faculty effectiveness through combined portfolios of faculty pedagogy and student learning. Looking at the quality of student work as it correlates to how information is presented could be a useful lens. Yet, one must consider how faculty would find time amid all other requirements to expend the effort needed to make this a truly meaningful exercise or merely acquiesce to the continued use of student opinions. After all—if nothing else—course evaluations are likely the most efficient way to gather this information. Moreover, any efforts to judge teaching through portfolios could lead to faculty feeling pressure to ensure only the materials going into a portfolio are truly mastered.

Course evaluations definitively matter for individual faculty members—especially those teaching as adjuncts on semester-to-semester contingencies or seeking tenure. But today they seemingly matter to higher education more generally. Student opinions can drive which faculty are brought back to continue teaching our ever increasingly diverse student populations. While how this data is collected unquestionably matters, so too does what we do with it.

If department chairs and other academic administrators are aiming for pure efficiency and merely examining a rank-ordered print out of faculty scores, concerns about bias should rightfully increase. Reading student comments, seeking patterns and holistically evaluating faculty leads to a stronger evaluation. The noise in student feedback will not lend itself to distinguishing the very strong teachers from the strong—but is that actually its intended purpose? Or is it about looking for consistently lower performing faculty in the classroom or concerning comments that merit a closer examination as part of a larger process?

Hopefully the above has shown that institutions should examine how student feedback is captured and used—which we readily admit is a considerable effort given the issue. But, if done properly, it can lead to a richer way of understanding what is and what is not working in our classrooms. To meaningfully do this, though, we need to be able to separate some of the myths and realities surrounding course evaluations. Above we discussed the lay of the land, but many conversations about course evaluations happen either without the presence of data or small sample sizes.

But data is available that would allow us to do much more—it is possible to separate some myths from realities. Below, we present and analyze six commonly held beliefs regarding course evaluations and attempt to assess the veracity of each.

Students who take course evaluations outside of class time are more likely to be critical in their comments and ratings

As course evaluations have more routinely moved to online platforms, faculty have raised concerns regarding the impact of students completing course evaluations outside of the classroom. Assumptions about students completing late at night while at home have led some to worry about how seriously students take evaluations when out of the classroom, or the impact of pressure from other students if taken in a group or social setting.

Comments on course evaluations only reflect the extremes and consequently are not constructive

Open-ended comments on course evaluations provide a separate set of datapoints for chairs, deans and provosts to consider when evaluating faculty. But a commonly held belief is that comments tend to come only from students who were highly satisfied or highly dissatisfied with a course. It is rare—in the opinion of many—to receive well-reasoned comments that reflect both areas of strength and weakness.

Course evaluation instruments do not accurately measure what faculty and administrators want

At the most basic level, concerns have been raised about the actual quality of course evaluation instruments in accurately measuring the information faculty and administrators would benefit from having. If questions are not well-vetted and lack shared meaning between faculty and students, the results will struggle with both validity and reliability.

Low response rates skew course evaluation results

As course evaluations move more toward online, a common faculty concern centers on response rates. If only thirty percent of students respond to an evaluation, can the data be deemed valid? What are the assumed attitudes of students who do not complete an evaluation? It is essential to understand how response rates interplay with the quality of the data ultimately produced.

Respondents have a consistent attitude across different evaluations

Many believe students are selective in which evaluations they complete or do not complete for myriad reasons. If faculty who believe students do not differentiate between different instructors or courses when filling out course evaluations—or only complete for faculty they particularly enjoyed or disliked—are correct, we could expect to see little to no variation in course evaluation results for these students.

Evaluation comments can be a predictor for average rating on course evaluations

In some cases, students may opt to forego any type of quantitative ranking and instead only offer comments in a text box. While this provides useful qualitative information for consideration, it does lead to wondering how the student's comment would correlate with numeric scores if they had been provided. Some go as far as to argue that the comments can actually predict faculty ratings on evaluations.

To examine the actual veracity of these beliefs, we utilized data gathered from 12 institutions of higher education in the United States that make use of the Campus Labs course evaluation system for conducting their student feedback process. Campus Labs routinely uses data collected from its partner institutions to provide landscape analyses and takes necessary steps to anonymize this data.

Regarding limitations, since this is pulled data from the Campus Labs system, we are unable to investigate any hypotheses related to course grades or gender and ethnicity biases, as those data points are not reliably made available to Campus Labs by the sampled campuses—this is an area in which we recognize the need for further industry study.

The institutions in this study were chosen to ensure adequate response counts, geographic dispersion and representation of various institution types to the extent possible. We intentionally selected institutions from across the United States and accredited by various regional accrediting bodies—as detailed in the accompanying side bar.

The data includes responses from July 1, 2016, onward in order to maintain recency and manage the total number of datapoints being examined. Information on the institutions and the number of course evaluation responses is included in the table below.

Selected Institutions

- 5 institutions from Southern Association of Colleges and Schools Commission on Colleges (SACSCOC)
- 4 institutions from Higher Learning Commission (HLC)
- 1 institution from Middle States Commission on Higher Education (MSCHE)
- 1 institution from New England Commission of Higher Education (NECHE)
- 1 institution from WASC Senior College and University Commission (WSCUC)

Institution	Two- or Four-Year Institution	Responses
1	Two-year	18,303
2	Two-year	61,058
3	Two-year	14,930
4	Two-year	25,611
5	Two-year	18,584
6	Two-year	67,729
7	Four-year	337,935
8	Four-year	408,898
9	Four-year	736,909
10	Four-year	90,156
11	Four-year	482,607
12	Four-year	39,546
Total		**2,302,266**

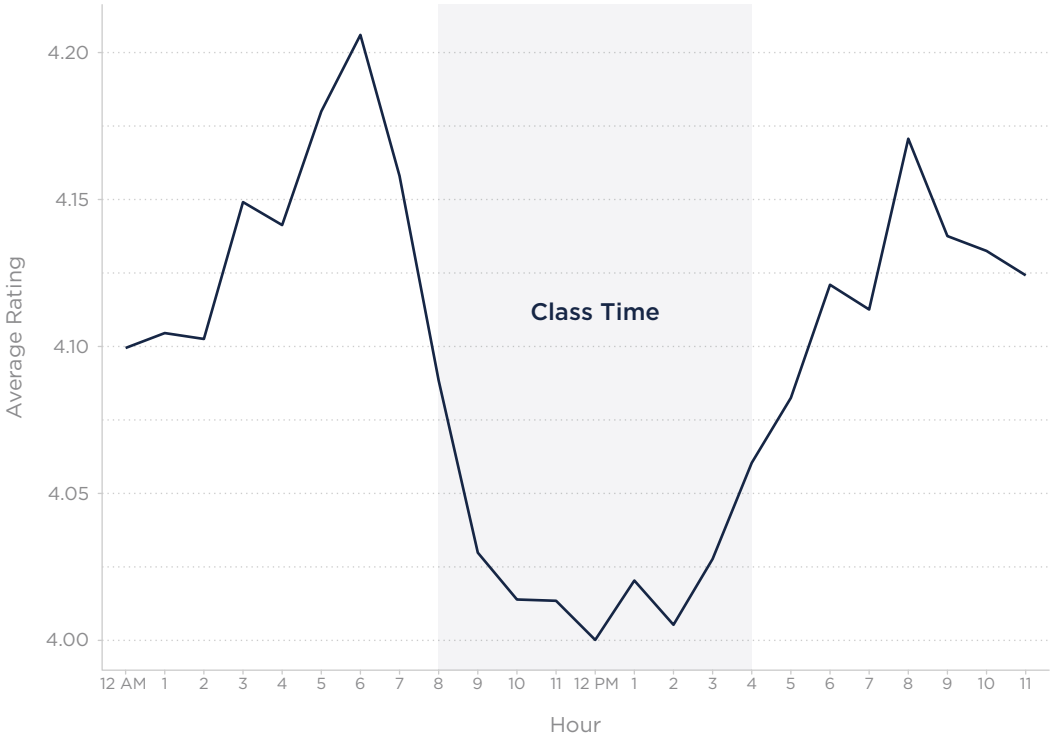
In total, more than 216,000 respondents provided the 2.3 million evaluation responses used to examine each belief. Below we discuss the data and methods used to analyze the course evaluation beliefs and attempt to rule on the veracity of each.

Students who take course evaluations outside of class time are more likely to be critical in their comments and ratings

To analyze this commonly held belief, we aggregated data for respondents at all examined institutions, averaging out student ratings after rescaling answers into a five-point scale and by hour in which they were completed. Given the geographic diversity of institutions represented, the time zones were all adjusted to Eastern time.

We arbitrarily set class time hours to between 8 a.m. and 4 p.m., estimating that the bulk of campus courses occur during this timeframe—we acknowledge that evening classes and some two-year institutions are more likely to have an increased number of classes after 4 p.m.

Figure 1: Evaluation Ratings by Time of Day



Contrary to what this belief suggests, Figure 1 shows that evaluations completed during class times are likely to be slightly more critical in average rating than out-of-class responses—and the trend is fairly steady. The highest average ratings, based on Figure 1, occur around 6 a.m. with the lowest happening at noon. In Figure 2, we see the division between two- and four-year institutions, time of day completed and overall ratings.

Results demonstrate the idea that students tend to be more critical when completing online evaluations outside of traditional class times is incorrect.

Figure 2: Evaluation Ratings by Time of Day Split by Institution Type

These results add a series of additional layers to consider. First, the split graphs show that overall, average ratings run approximately two-tenths of a point higher for our sampled two-year institutions compared to their four-year counterparts. Moreover, the

impact of time the evaluations were completed disappears for the two-year institutions. While four-year institutions show peaks outside of typical class times and a valley during, two-year institutions have a more consistent distribution.

Ultimately, results demonstrate the idea that students tend to be more critical when completing online evaluations outside of traditional class times is incorrect—and for four-year institutions, the opposite appears to be the case.

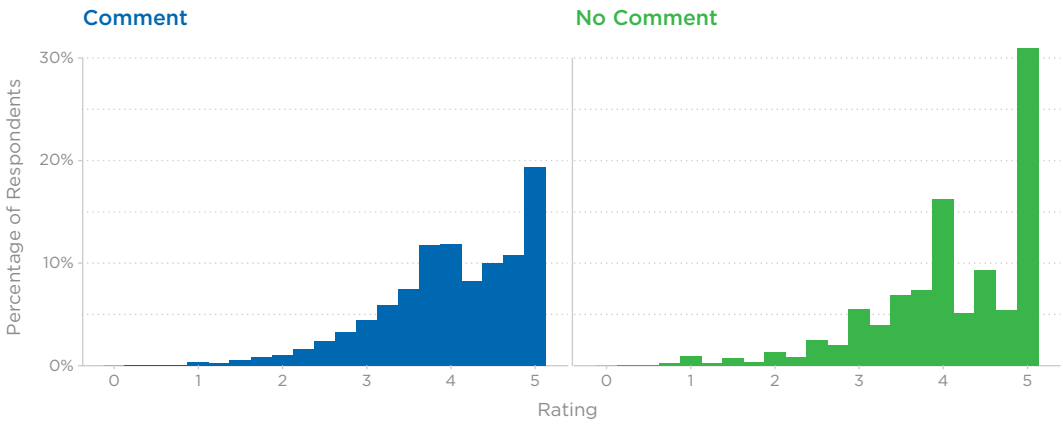
Comments on course evaluations only reflect the extremes and consequently are not constructive

For this belief, we created histograms to compare the average ratings of respondents who left comments to those who did not. Each pulls data from all 2,302,266 respondents, with 1,088,383 (47.3 percent) leaving comments. Individuals at two-year institutions left comments at a 74.8 percent rate while 44.6 percent of students at four-year institutions left comments. In Figure 3, we see the aggregated data for all institutions.

74.8%
of two-year students left comments

44.6%
of four-year students left comments

Figure 3: Distribution of Ratings Based on Comments

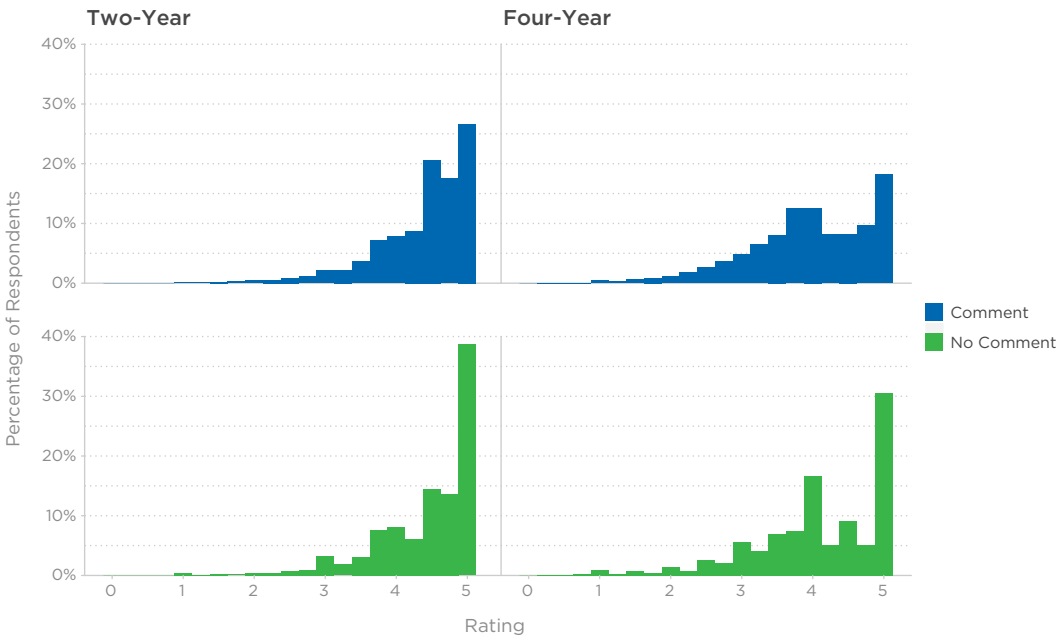


To begin, the results show a negative skew for both those who leave comments and those who do not, which suggests concerns about bimodal extremism being present in course evaluations is unfounded—particularly when considering qualitative comments offered. For the students that leave comments, a bulk of those rate faculty above the median value.

For the students that leave comments, a bulk of those rate faculty above the median value.

What stands out most is the percentage of students who rate faculty at the maximum end of the scale but do not leave any comments regarding why. It appears then that students offer a significant volume of comments for faculty they rate between a 3 and 5, suggesting they may be offering useful insights as opposed to simply singing praises or raising concerns.

Figure 4: Distribution of Ratings Based on Comments Split by Institution Type



In Figure 4, we again split the analysis between two- and four-year institutions to look for any notable patterns or differences. On the whole, the same patterns remain—respondents at two-year institutions tend to rate faculty higher, whether they leave comments or not. For this belief, there is no evidence to support a fear of receiving

only very negative or very positive feedback, which presents little value. However, this data does open a line of questioning on how we can better design instructor feedback systems to encourage more students to leave comments—along with helping to ensure provided comments are best utilized to assist faculty in better ensuring student success in subsequent semesters.

Course evaluation instruments do not accurately measure what faculty and administrators want

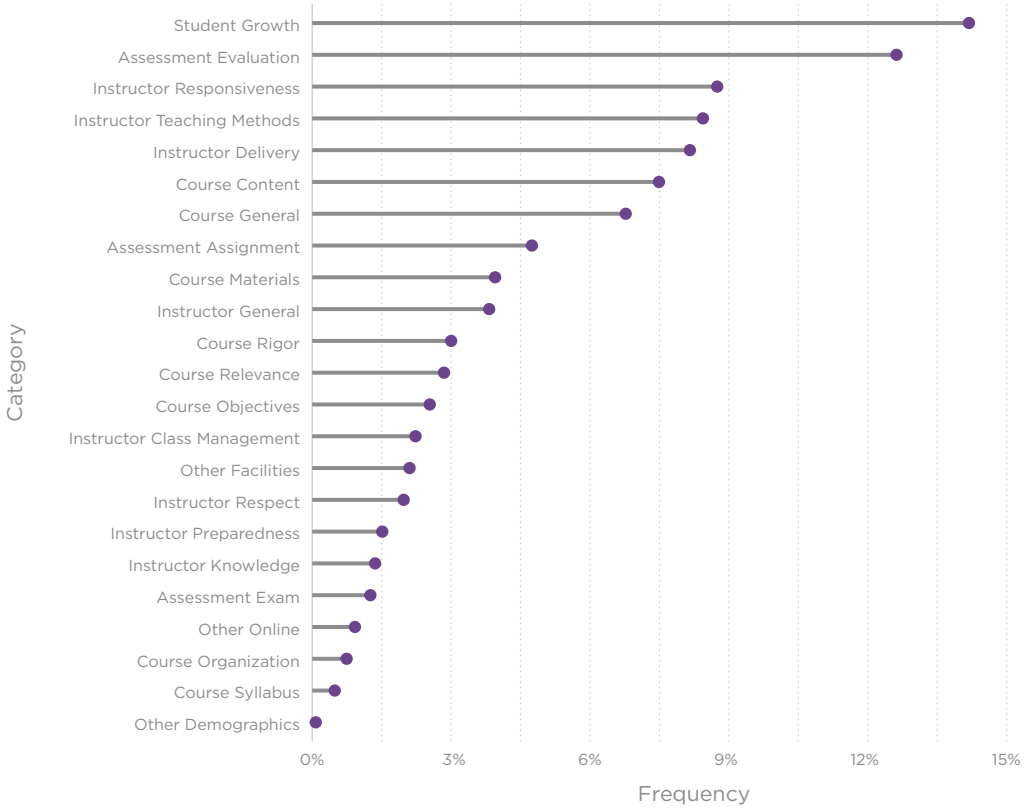
In order to examine this belief, we utilized our Campus Labs algorithm for classifying course evaluation questions—click here for more details[1]—to categorize the 3,898 various questions included in this analysis into 23 separate categories focused on student growth, assessment, instructor behaviors, course design and facilities. A raw distribution of categories is included here, while a lollipop graph visually shows the same as percentages in Figure 5.

Distribution of Question Categories

Category	Count
Student growth	553
Assessment evaluation	492
Instructor responsiveness	341
Instructor teaching methods	329
Instructor delivery	318
Course content	292
Course general	264
Assessment assignment	185
Course materials	154
Instructor general	149
Course rigor	117
Course relevance	111
Course objectives	99
Instructor class management	87
Other facilities	82
Instructor respect	77
Instructor preparedness	59
Instructor knowledge	53
Assessment exam	49
Other online	36
Course organization	29
Course syllabus	19
Other demographics	3

[1]Read more about the Campus Labs algorithm for classifying course evaluation questions at www.campusintelligence.com/blog/2017/05/24/what-are-you-learning-from-your-course-evaluations

Figure 5: Distribution of Question Categories



To give a sense of examples found within each category, the table below presents categories and a randomly pulled sample of an evaluation question found within it.

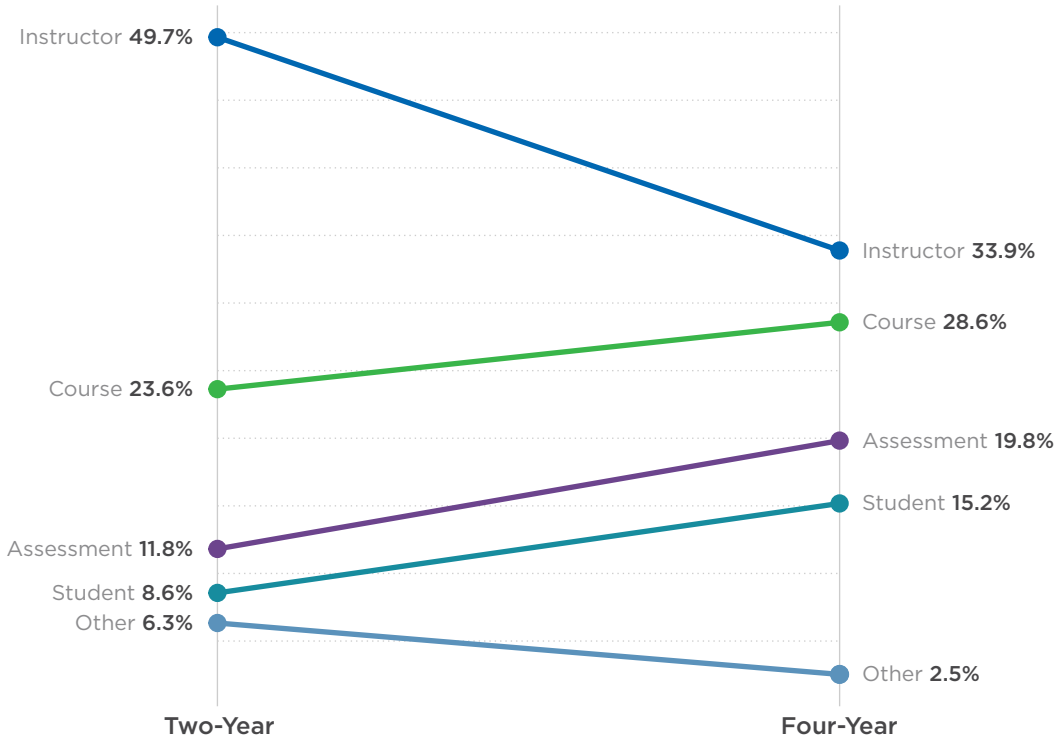
Category	Sample
Student growth	My instructor effectively challenged my thinking about the subject matter
Assessment evaluation	Evaluative and grading techniques (tests, papers, projects, etc.) were:
Instructor responsiveness	Provided timely feedback on graded work
Instructor teaching methods	The instructor encouraged class participation
Instructor delivery	My instructor's teaching strategies helped me to understand course content
Course content	Relationships among course topics are clearly explained
Course general	Overall, I would rate this course as:

Category	Sample
Assessment assignment	The projects assigned are consistent and reasonable
Course materials	Equipment used in the course supported course objectives
Instructor general	Rate the professor's effectiveness as an instructor
Course rigor	The course was sufficiently challenging
Course relevance	The instructor helped me to understand how the course fits my program of study
Course objectives	This course had clearly stated objectives
Instructor class management	My instructor met the class as scheduled in the syllabus
Other facilities	I can work (talk with tutors and write) comfortably in the Writing Center space
Instructor respect	The instructor dealt with all student perspectives respectfully
Instructor preparedness	My instructor seems well-prepared for class
Instructor knowledge	The lab instructor was knowledgeable about the course material
Assessment exam	Classwork and assignments prepare me to complete quizzes and exams
Other online	The technology tools were appropriate for the type of online course
Course organization	Course organization was:
Course syllabus	The course syllabus and schedule were clear and easy to follow
Other demographics	Please evaluate your room in the Residence Hall in terms of the overall cleanliness and size

While the lollipop graphic is useful to examine the results in aggregate, it is less useful when comparing two- and four-year institutions. Consequently, we have created slope graphs that directly compare the distributions for each question type. In the table below, we get a breakdown of the raw question counts for each overarching theme and institution level while in Figure 6 we have an aggregate slope graph showing how the question themes differ between the two institution types.

Category	Two-Year Count	Two-Year Percent	Four-Year Count	Four-Year Percent
Instructor	290	49.7%	1,123	33.9%
Course	138	23.6%	947	28.6%
Assessment	69	11.8%	657	19.8%
Student	50	8.6%	503	15.2%
Other	37	6.3%	84	2.5%
Total	**584**		**3,314**	

Figure 6: Aggregate Slope Growth by Institution Type



Two-year institutions are more directly interested in assessing the effectiveness of the instructor while four-year institutions more routinely emphasize courses, assessment and students.

Based on the types of questions asked on course evaluations, it appears two-year institutions are more directly interested in assessing the effectiveness of the instructor while four-year institutions more routinely emphasize courses, assessment and students. This could reflect the different missions and emphases of these institution types. In the table below, we break down the analysis to an additional layer by examining question counts per topic by thematic area for two- and four-year institutions.

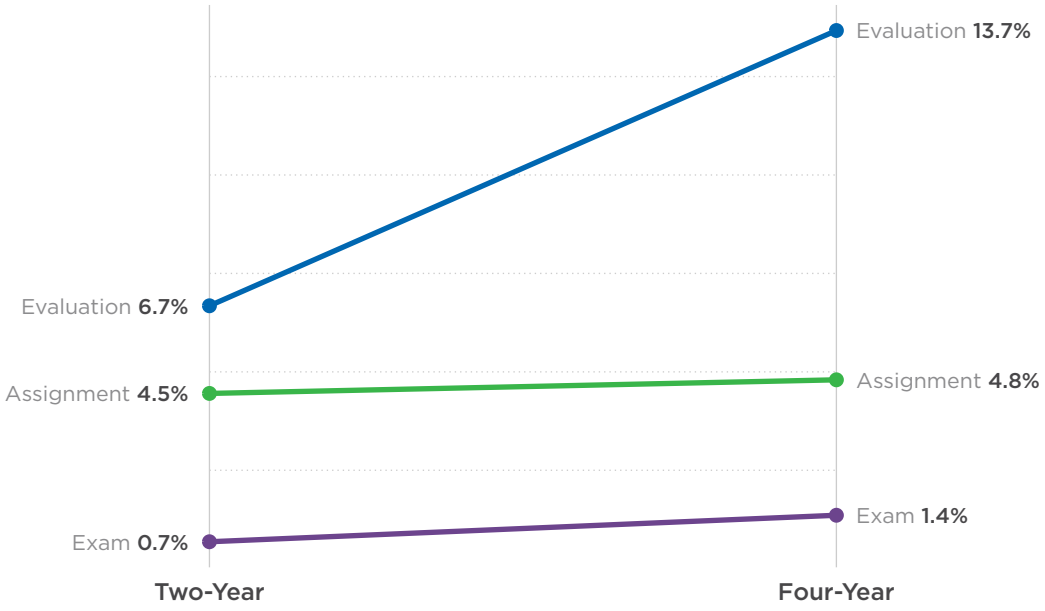
Question Count per Topic by Theme

Two-Year Institutions				Four-Year Institutions			
Topic	Theme	Count	Percent	Topic	Theme	Count	Percent
Assessment	Assignment	26	4.45	Assessment	Assignment	159	4.80
Assessment	Evaluation	39	6.68	Assessment	Evaluation	453	13.67
Assessment	Exam	4	0.68	Assessment	Exam	45	1.36
Course	Content	24	4.11	Course	Content	268	8.09
Course	General	21	3.60	Course	General	243	7.33
Course	Materials	29	4.97	Course	Materials	125	3.77
Course	Objectives	30	5.14	Course	Objectives	69	2.08
Course	Organization	2	0.34	Course	Organization	27	0.81
Course	Relevance	7	1.20	Course	Relevance	104	3.14
Course	Rigor	9	1.54	Course	Rigor	108	3.26
Course	Syllabus	16	2.74	Course	Syllabus	3	0.09
Instructor	Delivery	26	4.45	Instructor	Delivery	292	8.81
Instructor	General	33	5.65	Instructor	General	116	3.50
Instructor	Knowledge	32	5.48	Instructor	Knowledge	21	0.63
Instructor	Management	27	4.62	Instructor	Management	60	1.81
Instructor	Methods	49	8.39	Instructor	Methods	280	8.45

Two-Year Institutions				Four-Year Institutions			
Topic	Theme	Count	Percent	Topic	Theme	Count	Percent
Instructor	Preparedness	10	1.71	Instructor	Preparedness	49	1.48
Instructor	Respect	7	1.20	Instructor	Respect	70	2.11
Instructor	Responsiveness	106	18.15	Instructor	Responsiveness	235	7.09
Other	Demographics	0	0.00	Other	Demographics	3	0.09
Other	Facilities	17	2.91	Other	Facilities	65	1.96
Other	Online	20	3.42	Other	Online	16	0.48
Student	Growth	50	8.56	Student	Growth	503	15.18
Total		**584**				**3,314**	

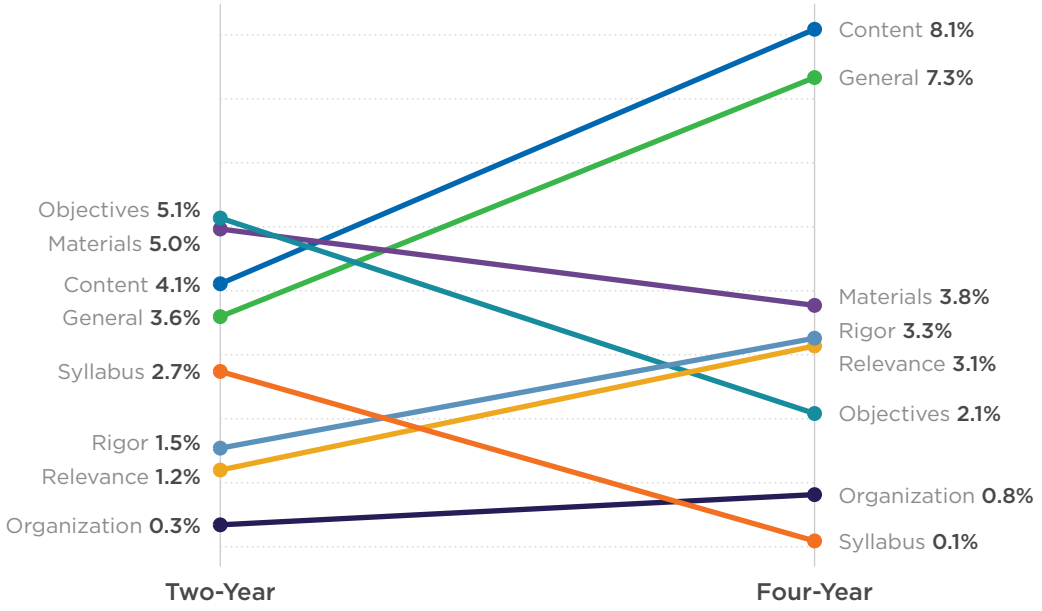
Figures 7 through 11 show similar, but more detailed, slope graphs with a breakdown for each theme.

Figure 7: Aggregate Slope Growth by Institution Type for Assessment



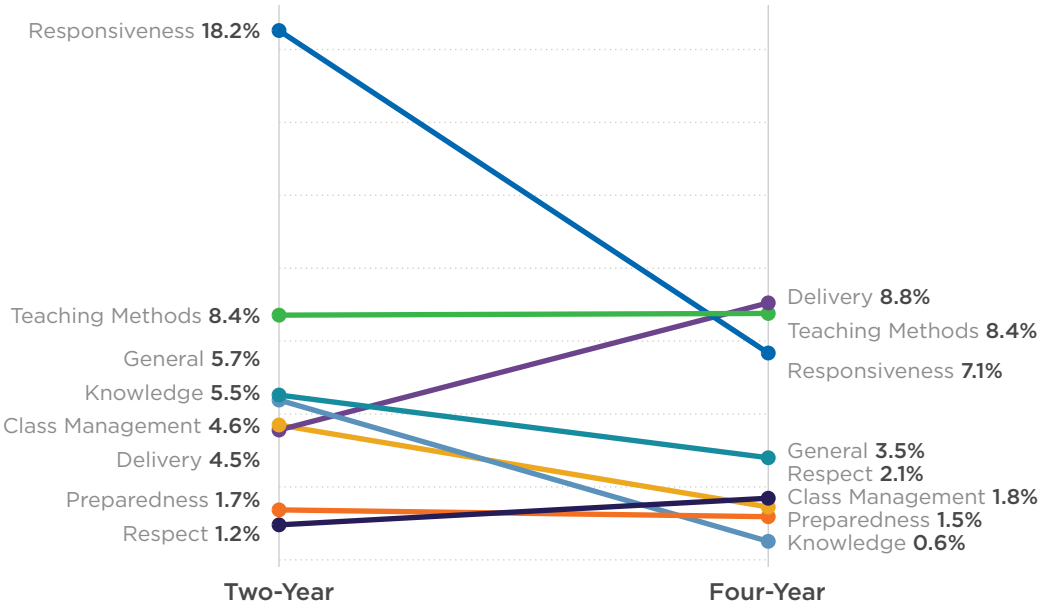
When looking at the theme of assessment, what stands out is the greater emphasis put on evaluation by four-year institutions.

Figure 8: Aggregate Slope Growth by Institution Type for Course



Within the course theme, four-year institutions ask more about content, rigor and relevance, while two-year institutions seem to stress objectives, materials and syllabi.

Figure 9: Aggregate Slope Growth by Institution Type for Instructor



The instructor theme suggests two-year institutions focus student feedback more on responsiveness, knowledge and class management when compared to their four-year counterparts—that instead devote more evaluation space to delivery.

Figure 10: Aggregate Slope Growth by Institution Type for Other

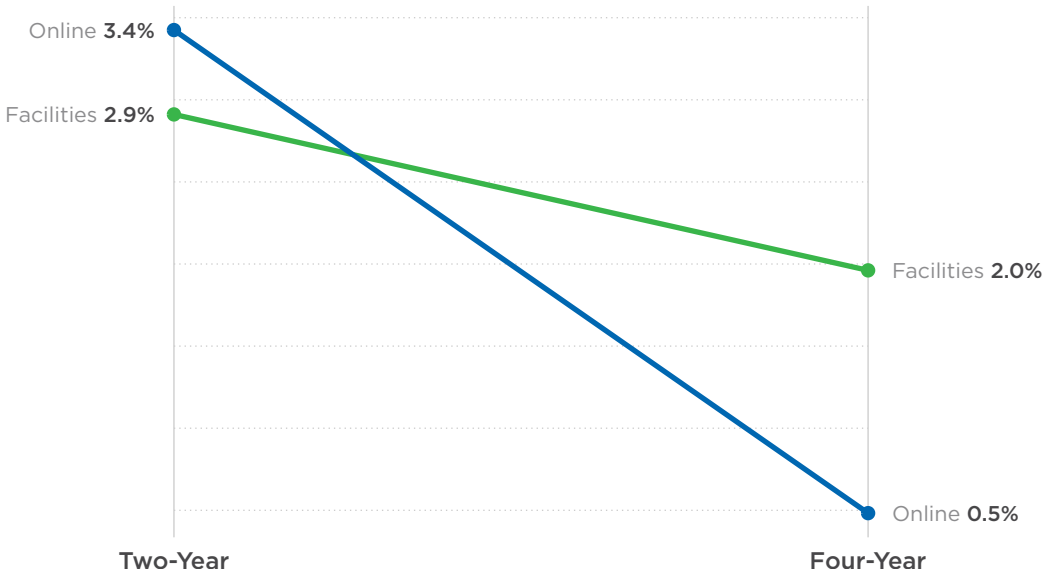
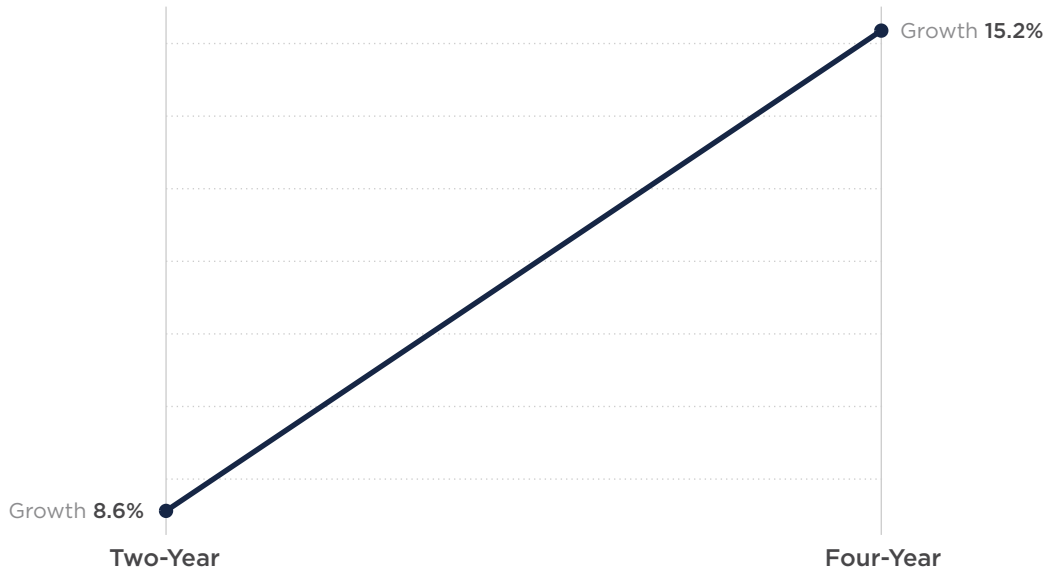


Figure 11: Aggregate Slope Growth by Institution Type for Student



There are measurable differences across several themes in how course evaluation instruments are constructed at institutions of various levels.

In terms of online and facilities, as seen in Figure 10, two-year institutions appear more likely to solicit student feedback—and as seen in Figure 11, four-year institutions focus significantly more on student growth.

What this data suggests is that there are measurable differences across several themes in how course evaluation instruments are constructed at institutions of various levels. Though, the unearthing of these various differences does not provide any evidence regarding this overarching belief. If different focal points emerge due to deliberate design choices by faculty and administrators at two- and four-year campuses, then the instruments very well could be measuring what faculty and administrators want. If, however, these discrepancies surface due to random chance, we may need to encourage greater intentionality in how evaluations are formulated. After all, it does not matter how many students respond if we are not asking meaningful questions that provide actionable data for faculty.

Low response rates skew course evaluation results

In order to analyze the veracity of this belief, we begin by creating a boxplot examining the average rating of courses taught by the same professor when response rates were low as compared to when response rates were high. To determine what counts as low and high response rates, we calculated the largest difference in response rate within a course—if there was greater than a 35 percent gap, the minimum and maximum response rates for that course were chosen.

Figure 12: Boxplot Comparing Average Ratings with Different Response Rates

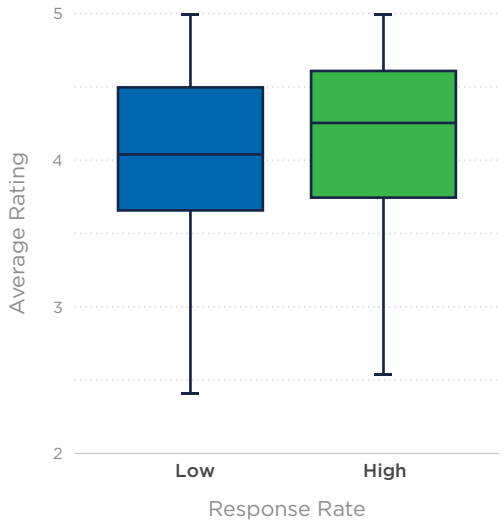
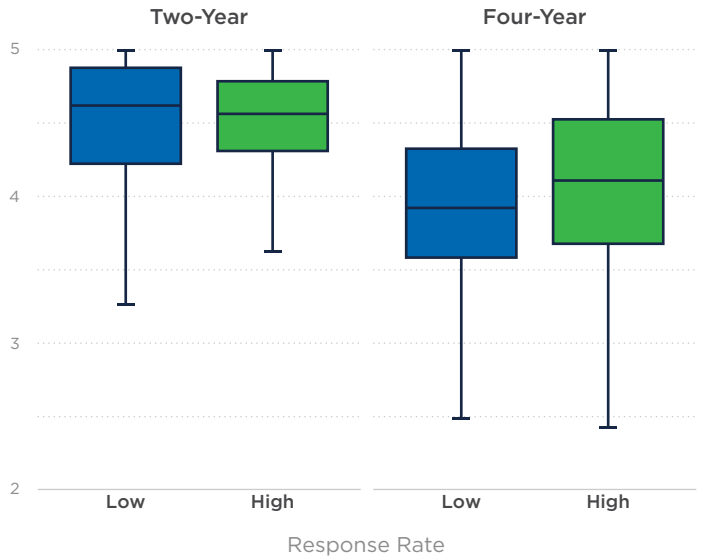


Figure 13: Boxplot Comparing Average Ratings with Different Response Rates Split by Institution Type



Any section with fewer than ten students enrolled or taught by multiple instructors was intentionally excluded. The boxplot in Figure 12 compares the averages for 8,308 sections representing 4,154 unique professor-course combinations.

The boxplot shows that while there is a slightly higher average rating for high response courses, it is not a substantively meaningful difference. The same diagram can be generated to show any potential division between two- and four-year institutions. Figure 13 uses the same data as the aggregate boxplot, with 1,713 course sections being from two-year institutions and 6,595 sections from four-year institutions.

There is less of a difference between low and high response rate average ratings at two-year institutions than at four-year institutions—but what little difference there is actually suggests courses with lower response rates have higher ratings at two-year institutions..

Beyond the boxplots, we also ran a linear model for this belief to determine whether response rates have an effect on ratings. The results show that there is actually a statistically significant difference, but it is a very small overall effect, at .13 points on a five-point scale. The response rate only explains 1.35 percent of the variance in course ratings. Ultimately, this belief is shown to be true when looking at statistical significance—but, concerns about response rates are substantively shown to be potentially overblown.

Linear Regression Model Results for Average Ratings and Response Rates

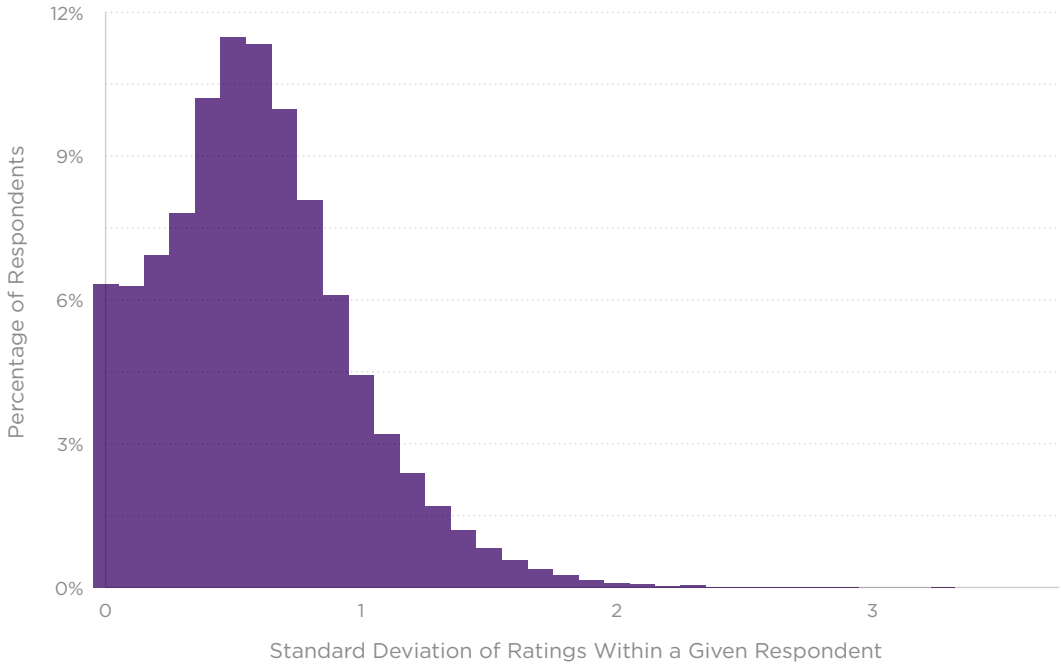
	Model 1: Min and Max Response Rates	Model 2: Level
Intercept	4.028	4.415
Max Group	0.133	0.127
	(10.660)*	(10.830)*
4 Year Level		-0.484
		(-33.31)*
Observations	8308	8308
F	113.700***	128.840*
Adjusted R2	0.013	0.130

**Significant of p < 0.001*

Respondents have a consistent attitude across different evaluations

To examine this belief, we produced a histogram showing the distribution of standard deviations between all the ratings a given respondent has submitted across different course evaluations. In total, the graph plots data from 191,755 students who completed more than one evaluation—Figure 14 shows the results.

Figure 14: Histogram Showing Distribution of Standard Deviations of Ratings within a Given Respondent



The distribution has a positive skew and suggests an overwhelming majority of students offer unique evaluations for different courses and instructors. Only slightly more than six percent of respondents have no standard deviation, showing that each rating they offered was identical. Figure 15 shows the same information as Figure 14, but it is separated between two-year and four-year institutions and is composed of data from 34,546 students that attended two-year institutions and 156,209 students that attended four-year institutions.

An overwhelming majority of students offer unique evaluations for different courses and instructors.

Figure 15: Histogram Showing Distribution of Standard Deviations of Ratings within a Given Respondent Split by Institution Type

The split histograms suggest students at two-year institution are more likely to provide similar ratings across multiple evaluations than their four-year counterparts. The distribution shows a stronger positive skew occurring at two-year institutions. Overall most students have ratings, on average, within one point of their mean rating. Ultimately, while the evidence shows the belief to be partially founded, it is less visible in four-year institutions.

Evaluation comments can be a predictor for average rating on course evaluations

To see whether this belief is supported by data, we measured the sentiment level of each comment, classifying them as positive or negative utilizing a dictionary-based approach to sentiment classification. This sentiment algorithm accounts for valence shifters (e.g., negations, adversative conjunctions). For example, the comment "The professor is always late and doesn't return papers in a timely manner" would be labeled as negative. Conversely, "Helped answer any questions that we had and did everything that they could to help students" would be labeled as positive.

We randomly split the data set (n = 1,068,437) into a training and test set using 80 and 20 percent of the data respectively. We then crafted a simple linear model to determine whether or not comments could in fact be used to predict course evaluation ratings. The model includes average sentiment of a comment, word count and the hour of the day that the comment was left. The results of the model are presented in the table below. We then applied the model to the test data set to compute error statistics. There were 854,750 comments in the training dataset used to build this model.

Linear Regression Model Results for Predicting Average Rating

	OLS Model
Intercept	3.791
Average Sentiment	.792
	(288.520)*
Word Count	-.001
	(-100.690)*
Hour of Day	.006
	(32.780)*
Observations	854,750
F	34550*
Adjusted R^2	0.108

**Significant of p < 0.001*

The overall model explains 10.75 percent of the variance in average ratings. The average sentiment has the strongest predictive power, although its statistical significance would be classified as small, while both word count and hour completed are also statistically significant. Ultimately, the more positive the sentiment, the fewer the words, and the later in the day comments are made all contribute to higher average ratings.

Contributing factors to higher average ratings include: more positive sentiment, fewer words and the later in the day comments are made.

We then applied the model to the test data (n = 213,687) set to understand how well the model would work to predict average ratings. We computed the residuals for actual values in comparison to the predicted value—a residual is computed by taking the predicted rating and subtracting it from the actual average rating (residual = actual – predicted). For example, if our model predicted a rating of 4.6 but the actual student's average rating was 3.6 the residual for that student would be 3.6 – 4.6 = -1.

Figure 16: Residuals from Linear Regression Model Results for Predicting Average Rating

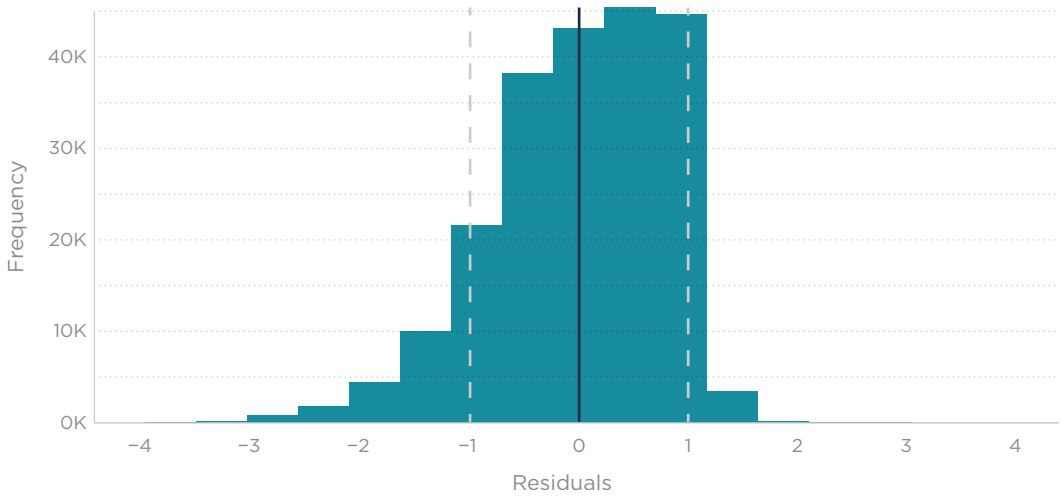


Figure 16 shows the distribution of how far away from the actual average rating that the model was. A value of 0 (blue line) indicates a perfect prediction. A value between -1 and 1 (grey dotted lines) indicates that the model prediction was within one point of the actual average rating. Generally, the model was within one point on a five-point scale for predicting average rating on data it had not seen before. Root-mean-square error (RMSE) is a metric that is used to distill these residuals into an interpretable metric. Loosely, RMSE can be thought of as the average of how far off from reality the model predicts. If the model is predicting perfectly, this value will approach zero. RMSE is in the same units as the data, so we can think of this number as being how many points (on a five-point scale), on average, the model is away from the correct average rating.

Our model performed reasonably well, with a RMSE of .79 points, that is, we can expect on average that the model will be off from the actual rating by 8/10 of a point. This suggests that this specific belief can be verified by data available from our sampled institutions, but there is the caveat of how substantively important the power of the finding is.

Ultimately, campuses need to have data-informed discussions surrounding course evaluations.

Conclusion and suggested areas for further research

The analyses for each commonly held belief in this paper hopefully show the possibilities and power of examining course evaluation data as more than a point-in-time mechanism for evaluating faculty effectiveness in the classroom. The data used from our sample institutions should be available on any campus in order to duplicate the analyses and for the identification of campus-specific intricacies.

Moreover, individual campuses have the data to take the analyses a step further and examine how grades, gender and other demographic factors impact evaluation ratings. Basic, direct analyses will allow for a campus to see if there is a bias at play and identify patterns in how some students choose to rate a faculty member low. With our anonymized data from the campuses used in this particular study, we were unable to do so, even if we know those are some of the more interesting—and for many, most pressing—beliefs about course evaluations.

Ultimately, campuses need to have data-informed discussions surrounding course evaluations. Why are they conducted? How are they being used by chairs, deans and provosts? What value does a student receive by taking the time to complete them? And most importantly, what concerns do faculty have about evaluation instruments, processes and the use of results? It is then up to administrators and faculty to together examine if concerns are well-founded based on the data—or, if further investment is needed to educate all internal stakeholders on the difference between course evaluation myths and realities.

Campus Labs Data Science

The Campus Labs Data Science Team has the privilege and a shared responsibility to empower institutions to make impactful changes through the strategic use of data—we accomplish this by understanding the interconnected interactions of students, families, faculty and staff within a learning community. This complex network of people, places and events generates rich stores of data that can be harnessed and modelled to understand and act in ways that bring success. As such, we are committed to protecting the quality of data, best in class data modeling and presentation of continually improving results.

The quality of analysis is first contingent upon the quality of data. We are advocates of careful, responsible collection of relevant variables that are used to enrich the lives of all our stakeholders. We partner with campuses to improve the accuracy and completeness of their data. Diligence in improving data quality provides our modeling techniques with greater signal while reducing noise.

The Data Science Team are life-long learners and use current analysis methods to provide an actionable representation of the complexity of campus life. These techniques can be used to understand not only traditional, quantitative data, but also the rich, complementary qualitative data—providing realistic summarizations of data that are presented back to our stakeholders in actionable ways.

These summary models are continually updated to reflect new information that is collected. The results may show up in many different forms, all of which empower stakeholders to make informed decisions. This analysis results in new graphics, widgets, variables, reports and other features—but, the true impact our team has is in the way data, analysis, and results equip students, families, and faculty to make decisions that equal success.

Tyler Rinker, Ph.D.
Manager, Data Science
Campus Labs

About the authors

Will Miller, Ph.D.
Executive Director, Institutional Analytics, Effectiveness and Strategic Planning
Jacksonville University

Will Miller, Ph.D., is executive director of institutional analytics, effectiveness and strategic planning at Jacksonville University. Most recently, Miller served as assistant vice president of campus adoption at Campus Labs. During his time at Campus Labs, Miller leveraged data best practices to help campuses across the globe make strategic, data-informed decisions. Prior to Campus Labs, he served four years as a faculty member, senior administrator and the SACSCOC liaison at Flagler College, where he oversaw the campus-wide outcomes assessment process, as well as planning and institutional research activities. Prior to joining Flagler, he held faculty positions at Southeast Missouri State University, Notre Dame College and Ohio University. He earned his master's degree in applied politics from the Ray C. Bliss Institute at The University of Akron, where he also earned his doctorate in urban studies and public affairs. He holds both a master's degree in political science and a bachelor's degree from Ohio University

Tyler Rinker, Ph.D.
Manager, Data Science
Campus Labs

Tyler Rinker, Ph.D., joined Campus Labs in 2015, where he leads the data science team in the use of data to understand the interconnected interactions of students, families, faculty and staff within learning communities. He also teaches courses at the State University of New York at Buffalo (UB) and develops open source analysis tools. Prior to this Rinker was an elementary and middle school educator. He completed his doctorate in curriculum, instruction and the science of learning at UB with a certificate of advanced study (CAS) in statistics, where he developed methods to investigate the intersections of people, language, action and environment. He holds a CAS in educational leadership and a master's degree in childhood and early childhood curriculum and instruction from Buffalo State College.

campuslabs

About Us

Campus Labs was founded to empower educational institutions to evolve in a data-centric world. Uncover a platform of integrated tools that drive an institutional mindset for insightful data connections.

The holistic Campus Labs framework includes solutions for assessment, retention and success, teaching and learning, student engagement, skills and achievement, and institutional effectiveness. Proudly serving more than 1,400 member campuses, discover more at **www.campuslabs.com**.

Media Inquiries

Contact Kyle Gunnels, Assistant Vice President of Communications
Email: **kgunnels@campuslabs.com**