5-10-2023

Distributional properties of the statistic of online student evaluations the mean does not mean what you think it means

Ben O. Smith

Dustin R. White

Jamie Frances Wagner

Patricia C. Kuzyk

Alex Prera

Distributional properties of the statistic of online student evaluations the mean does not mean what you think it means

Ben O. Smith [a], Dustin R. White[a], Jamie Wagner[a], Patricia Kuzyk[b] and Alex Prera[b]

[a]University of Nebraska at Omaha, Omaha, NE, USA;

[b]Washington State University, Pullman, WA, USA

To cite this article: Ben O. Smith, Dustin R. White, Jamie Wagner, Patricia Kuzyk & Alex Prera (2023): Distributional properties of the statistic of online student evaluations the mean does not mean what you think it means, Studies in Higher Education, DOI: https://doi.org/10.1080/03075079.2023.2211079

ABSTRACT

Student Evaluations of Teaching (SETs) are an integral part of evaluating course outcomes. They are routinely used to evaluate teaching quality for the purposes of reappointment, promotion, and tenure (RPT), annual review, and the rehiring of adjunct faculty and lecturers. These evaluations are often based almost entirely on the mean or proportion of the ordinal overall score with no regard to statistical noise. This study examines the distribution of the statistic (mean or proportion) when SETs are administered online and in-person. Using non-parametric procedures, we show that the size of the 95% confidence interval of the statistic is a function of response rates. Prior to COVID-19, online administration of SETs resulted in significantly more uncertainty than in- person administration because the in-person response rates were higher. Due to a decrease in in-person response rates in the post-COVID vaccine period, both methods result in significant levels of uncertainty of the true statistic value. In classes of fewer than 30 students, the 95% confidence interval of the statistic is wide enough for instructors to be considered for a teaching award in one semester or below average in another semester, while holding teaching quality constant.

KEYWORDS

Student evaluations; survey administration mediums; bootstrap; simulations; online teaching evaluation

Introduction

Few topics have the ability to unite nearly all members of academia, but a belief that Student Evaluations of Teaching (SETs) are a poor measure of teaching ability is one of them. Faculty need feedback about their classes, department chairs want to know how an instructor is doing in the classroom, and administrators want to assess faculty performance. All have turned to SETs, whether willingly or not. More recently, colleges and universities have provided the SET form online as a way to expedite the evaluation process: instructors do not have to give up valuable class time to administer the SET, online administration is quick and inexpensive, and SETs are a quantifiable way to measure students' opinions and perceived learning. The popularity of SETs leads to important questions. How do the in-person or online SET distributions differ? Put differently, do the same students fill out the SET regardless of method of delivery? If the students are different, how do they differ in their responses to the SET question set? This paper examines the distributions of the statistic (both the mean and proportion measure) of in-person and online SETs using non-parametric procedures. This is examined both in a period before COVID and a period after COVID vaccines were widely available.

Becker and Watts (1999) found that SETs are widely used in economics departments (and in many others) as the main way of evaluating teaching. In a follow-up study Becker, Bosshardt, and Watts (2012) found that 35% of departments were offering SETs online; only two department chairs from the 1999 survey responded that they used electronic SETs. This number has without a doubt increased due to the ease of use and cost advantages of online evaluations. As the movement from in-person to online SETs progressed, researchers began focusing on whether or not instructors received lower evaluations using the online administration of SETs.

Research (Donovan, Mader, and Shinsky, 2010, Heath, Lawyer, and Rasmussen, 2007, Fike, Doyle, and Connelly, 2010, Guder and Malliaris, 2010, Breitbach, Sankaran, and Wagner, 2016) has found that there were no substantial central tendency differences between the online and in-person SETs. In a follow-up study Sankaran, Wagner, and Breitbach (2018) found that the variance is much larger for online SETs compared to the in-person SETs. Other studies have also found evidence that different students are filling out online SETs compared to in-person SETs, and that this difference can alter the distribution of results (Dommeyer et al., 2002, 2004, Gamliel and Davidovitz, 2005, Morrison, 2013).

Closely related to our work, He and Freeman (2021) conduct Monte Carlo simulations based on administrative data at a medium-sized public institution of higher education. They find that response rates are often problematic for generating useful information about students' true assessments of courses. As future work, they suggest exploring similar modeling at other schools, evaluating the uncertainty in diverse questions asked of students, and using metrics beyond the arithmetic mean of evaluation scores (measured on a scale of 1-5). We extend their work by collecting data at a large public research university, simulating responses using non-parametric bootstrapping techniques, and by considering multiple aggregate measures of SET scores. Our findings reinforce and extend their conclusions, indicating that the noise inherent in SET scores at the course level is such that typical class sizes lead to a high risk of SET-derived measures deviating far from the true student evaluation value given typical response rates.

Understanding the distribution of the statistic (mean or proportion) of SETs is especially valuable when many universities, colleges, and departments rely on the information to make major decisions regarding tenure and promotion. This becomes even more important for non-tenure track positions such as a lecturer, instructor, adjuncts, or clinical professors whose contracts may be dependent upon teaching evaluation scores alone. This paper uses data from a large state university to compare the distribution of the statistic between traditional in-person SETs to the distribution of the statistic when administered online. Differential response rates are the driving force in this paper. In the pre-COVID period, we find substantial and statistically significant differences in the width of the 95% confidence interval at every class size. In this period, in-person SETs exhibit much less noise than online SETs. Despite this pattern, class sizes of less than 30 exhibit an interval so large using either method that an instructor could be considered wonderful or mediocre due simply to statistical noise and sampling. In the post-COVID vaccine period, response rates for the in-person administration dropped precipitously. Thus, while the difference in distributional proper- ties of the two administration methods declined in the post-COVID vaccine period, this is only because the in-person administration became less robust. In the current environment, either administration method could result in radically different scores for a single instructor due to randomness alone. This suggests that an observed SET average score, often used to evaluate faculty, contains significantly less information than many evaluators appear to believe.

Method

The difference in distributional properties of small-class SETs administered online and in-person can be determined using two different methods. First, online and in-person SETs could be administered in classes of various sizes. Alternatively, one could administer SETs online and in-person to a set of very large classes and use statistical techniques to simulate smaller classes of any size. This study uses the latter option. The advantage of this approach is that true underlying teaching is identical; the disadvantage is that the difference in response rate (when administered online versus in-person) is assumed not to be a function of class size.

Our study was conducted at a large research university. Toward the end of the semester, a physical SET was administered to the class by a member of the research team. This occurred at the midpoint of the online evaluation window, and the research member was not the instructor of the course. The physical SET was identical in every way to the online evaluation. Students were asked to fill out the physical SET in class. Students were informed that the physical SET was for a different purpose and thus they still needed to fill out the official online SET provided by the university. Our goal with this procedure was to replicate the selection process had the university administered an in-person SET; the university was already administering an online SET to the same class. This allowed us to compare the impact of the administration method on the distribution of the statistic while holding the actual underlying teaching constant. Thus, by design, some students filled out the physical SETs, some filled out the online SETs, and many filled out both.

This process was conducted in a total of four principles-level courses during the pre-COVID period: three in Fall 2018, one in Spring 2019. Three additional sections were included from Fall 2021 when the university had returned to in-person classes. The enrollment and SET response rate for each course included in the data can be seen in Table 1. We were concerned that the in- person administration might impact the online response rate. Thus, we compared the pre-pandemic online response rates (C.# 1–4 in Table 1) to the instructors' historical average response rate. Prior to the study, microeconomics sections taught by instructor A have averaged a 49% response rate (online). Similarly, macroeconomics sections taught by instructor A have averaged a 50% response rate (online). Instructor B's macroeconomics sections have averaged a 64% response rate (online). These historic results are noisy (low of 38%; high of 74%). However, there does not appear to be evidence that our in-person treatment is impacting the online evaluation response rate.

Note the response rate in the pre-COVID period versus the post-COVID vaccine period. Response rates are the driving factor of the distributions presented in the results section of this paper. In the pre-COVID period, the in-person response rate was always significantly higher than online. However, this changed in the post-COVID vaccine period. Unfortunately it did not change by improving the online response rate. Instead it decreased the in-person response rate.

In this study, we will focus on two questions: 'What is your overall rating of the instructor [First] [Last] in this course?' (instructor overall), and 'What is your overall rating of this course?' (course overall). Both question responses are provided on a five point Likert scale from 'poor' to 'outstanding,' and the scale is commonly converted to a 1-5 scale for the purposes of averaging. We focus on these questions as nearly every university has some version of the instructor and course overall questions. These questions also tend to be the questions emphasized by administrators when determining who is a 'good' or 'bad' teacher.

Selection

The differing selection process by students will drive the methodology presented in this study. The students who choose to fill out a student evaluation online fundamentally differ from those who would have filled out in-person evaluations. The selection of those who choose to fill out an in- person evaluation can be described as follows:

$$U\ (\text{Class Attendance}_{t}) \succ U\ (\text{Opportunity Cost}_{t}) \tag{1}$$

Table 1. Class size and response rate by course.

	C.#	Class size	Response rate		Instructor	Semester
			Online	In-person		
Microeconomics	1	236	50%	78%	A	Fall 2018
Macroeconomics	2	165	52%	70%	A	Fall 2018
Macroeconomics	3	251	51%	80%	B	Fall 2018
Macroeconomics	4	269	62%	77%	B	Spring 2019
Microeconomics	5	305	53%	32%	B	Fall 2021
Macroeconomics	6	137	48%	31%	B	Fall 2021
Macroeconomics	7	274	30%	32%	A	Fall 2021

‘C.#’ is the course number. In the main body of this paper, we will analyze courses 1 and 5. Other courses in the pre-pandemic period produce similar results to Course 1 while other courses in the post-COVID vaccine period produced similar results to Course 5. The results from the other courses are included in the appendix.

Where U(Class Attendance$_{t}$) is the utility the student received from attending class on day t, and U(Opportunity Cost$_{t}$) is the utility of next best option at that time period. Fundamentally, the student must only be self interested in improving their learning/grade for them to show up for class. We assume that, once an evaluation is revealed to the class, few students would choose to not fill it out given social pressure to fill out the evaluation along with the rest of the class when in attendance.

Notably, the U(Class Attendance$_{t}$) differs significantly between the pre-COVID period and the post- COVID vaccine period. Before COVID, missing class meant the student would have to ask a classmate for notes, which is a poor substitute for attendance. In the post-COVID vaccine period, video recordings were available allowing students to obtain much of the benefit of class attendance. There also exists a risk of COVID exposure that simply did not exist in the pre-COVID period. Overall, U(Class Attendance$_t$) appeared to decrease in the post-COVID vaccine period in comparison to the pre-COVID period. This is in concordance with the reduced in-person response rate during the post-COVID vaccine period.

The selection process for the online evaluations differs from the in-class process, and can be described using Equation (2):

$$U\ (\text{Evaluation}_j) > U\ (\text{Opportunity Cost}_j) \qquad (2)$$

Where U(Evaluation$_j$) is the utility that the student receives from filling out the evaluation itself (U(Opportunity Cost$_j$) is the utility of the next best alternative). For U(Evaluation$_j$) to be any value other than zero, the student must feel that evaluating the instructor benefits them in some way (this can include altruism and warm glow effects). It is possible that those who have a strong opinion (positive or negative) are therefore more likely to fill out the evaluation in this context than a student with less pronounced opinions of the course or instructor. This explanation is in line with the data presented by Sankaran, Wagner, and Breitbach (2018).

Simulating smaller classes

Our process of simulating smaller classes is essentially a bootstrapping procedure (Efron and Tibshir- ani, 1994). The underlying concept of a bootstrap is that the properties of the population can be modeled through re-sampling with replacement. Usually a re-sample size is chosen such that it equals the overall sample size. This reflects the total information in the sample thus correctly reflects the distribution of the statistic. Smaller classes, however, do not contain the information of a larger class. The amount of information in a smaller class of size c can be simulated by re- sampling with replacement c students for each repetition. This property of a bootstrap inspires our approach. Our simulation works as follows:

1. From sample *i*, *c students* are randomly selected with replacement (where *c* is the simulated class size, and *i* is the index number of the sample)
2. Of the randomly selected students, those who did not fill out the given SET instrument (or method of administration) are dropped
3. The remaining students are used to calculate the statistic for the given SET instrument
4. The procedure is repeated 10,000 times for each class size *c*

This procedure produces the distribution of the *statistic* of interest given a class size. For instance, if the statistic is the mean score and the class size is thirty, the resulting distribution is of the mean outcomes that the instructor could expect given the same underlying teaching.

In this study we calculate two statistics: the mean and the percent of responses that are a '4' or better. Calculating the mean of an ordinal variable is statistically incorrect. This assumes equal spacing of the values on the scale (e.g. the difference between 'poor' and 'below average' is the same as 'below average' and 'average'). Nonetheless, this statistic appears to be the most common measurement in practice. It is therefore important to understand the statistical properties of the measurement. Calculating a proportion above a chosen score level is a more statistically correct approach as suggested by McCullough and Radson (2011).

We compare the distribution of the statistic using multiple techniques. First, for simulated class sizes of 30, we use the two sample Kolmogorov-Smirnov test of the whole distribution (Conover, 1998); this test makes no assumptions regarding the distributions. Using this test we compare the online administration of the evaluations to in-person evaluations and determine if the distributions of the statistic are statistically different (the null hypothesis of the test is that they are the same). More importantly, we use Kernel Density Estimations (Scott, 2015) to visually show the differences in the online versus in-person distributions.[1]

Finally, we simulate the distributions of the statistic at increasing class sizes from 10 to 150 students. We extract the width of the 95% confidence interval for each of these simulations and fit a Locally Weighed Scatterplot Smoothing (LOWESS) curve (Cleveland and Devlin, 1988) to the set of 95% confidence interval observations for both the in-person and online evaluations. We have selected a class size range from 10 to 150 students as few administrators would trust an SET score from a class of fewer than 10 students and the confidence intervals contract substantially by 150 students.

Results

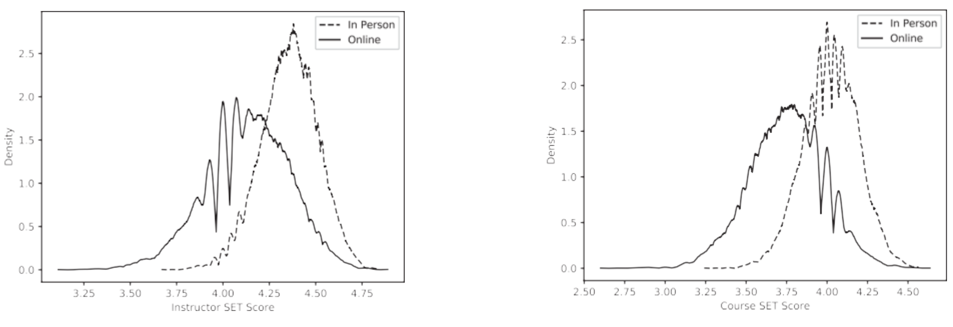
The results of the simulations described in the method section have the benefit of being easily rep- resented visually. Accordingly, we will present the results as figures and discuss the features of these figures below. While we will select figures that readily demonstrate the key lessons from our simulations (using Course 1 from Table 1), our results hold across *all simulations from all course sections in the pre-COVID period*. Visuals representing all other courses can be found in the appendix.

Results from Course 1 (pre-COVID period)

Mean SET scores

While statistically incorrect, it is common practice to use mean SET scores as a measure of teaching outcomes. Figure 1 presents the Kernel Density Estimation (KDE) of the mean SET score in simulated classes as the statistic of interest; Figure 1(a) presents the distributions of the instructor overall score and Figure 1(b) presents the distributions of the course overall score. The in-person and online data is independently analyzed and projected onto the same graphs for comparison purposes.

Figure 1. Course 1: KDE for mean SET scores in simulated classes of 30 students. KS test *p*-values = 0.000.



(a) Instructor

(b) Course

Two features distinguish the online scores from those given by students in-person: the width (variance) of the distributions and the modes. Looking at Figure 1, we observe thicker tails to the distributions of the mean SET scores when administered in an online format compared to the in-person format. This indicates that simply requesting SET scores through an online portal can significantly reduce the confidence that we assign to the mean measures of teaching outcomes even holding instructor and course constant.

The mode of the distributions are also noticeably different between the two delivery methods: as described above, a different group of students will likely complete SETs online and in-person. Thus, the position and/or height of the mode can vary. We conducted Kolmogorov-Smirnov (KS) distributional tests to compare the distribution of mean SET scores. Using these tests, we can reject the null hypothesis that the distributions are the same at *at least* the 99.9% confidence level for *all simulations across all courses and class sizes in the pre- COVID period*. The data collected by in-person and online SETs are distinct measures with distinct distributions.

The weakness of mean SET scores as an ability to accurately measure instructor quality becomes vividly apparent in Figure 2. Figure 2(a) presents the width of the 95% confidence interval for the instructor mean SETs at various class sizes while Figure 2(b) presents the width of the 95% confidence interval for the course SETs at various class sizes.

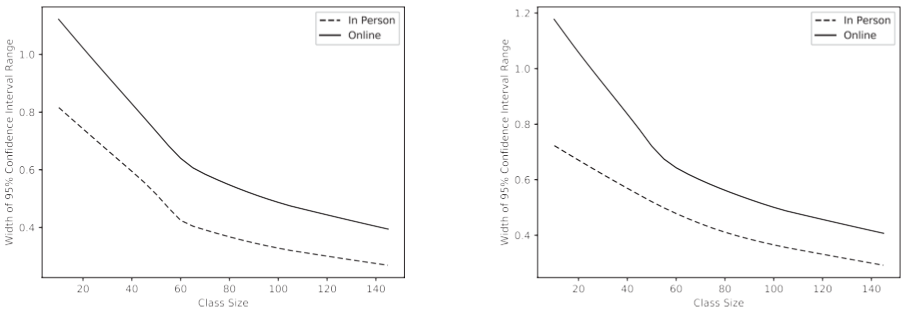
At a class size of 30, the 95% confidence intervals have a width of approximately 1.0 for online ratings. This means that, at this class size, an instructor could be rated at 4.5, but reasonably have a true statistic of 5.0 or lower than 4.0. Instructor or course ratings can be the deciding factor in outcomes such as promotion or retention, but the noise in this measure is large enough that little faith can be put in observed outcomes for even moderate class sizes.

Percent of ratings above 4 as a statistic

The more appropriate statistical measure of teaching outcomes using SET scores is to report the per- centage of responses that rate a measure at or above a given level. We choose to measure the pro- portion of responses that rate instructors (or courses) at or above a 4 on the 5-point Likert scale. The results of sampling from the distribution of responses on this statistic when assessing the instructor can be seen in Figure 3.

When considering the proportion of responses above a given threshold (an improved statistical measure), we see divergences between the online and in-person SET distributions similar to those observed in the mean response distributions. The tails of the online rating are much thicker, and the center of the distribution is noticeably lower for the online SETs.

Figure 2. Course 1: LOWESS curve for mean SET score in simulated classes of varying size. KS test *p*-values = 0.000.



(a) Instructor

(b) Course

Again, as in the case of the mean SET score, the KS test is significant at the 99.9% level, indicating that the distributions are distinct. The test maintains the same level of significance at all class sizes and across all instructors.
The confidence interval for the proportional statistic is so large that it should inspire no confidence in the measure’s ability to provide information regarding instructional outcomes. Figure 4 shows the LOWESS Curves for the proportion of online and in-person SET scores of a 4 or better.

Again, focusing on a class size of 30 students, the 95% confidence interval for the proportional statistic has a width of about 0.45 in online evaluations. In other words, one in twenty instructors who have a true statistic of 0.78 will have an observed statistic of 0.55 or lower or of 1.00. Five percent of instructors who receive an underwhelming 0.60 will have a true statistic of above 0.82 or below 0.38.

This result provides further reason to mistrust the results of SET scores (regardless of the method of administration), whether the score is measured as a mean value or as the proportion of responses above a given threshold. There is simply too much noise present in the evaluations to be able to assert that SET scores have any meaningful relationship with the outcomes that they attempt to measure.

Results from Course 5 (post-COVID vaccine period)

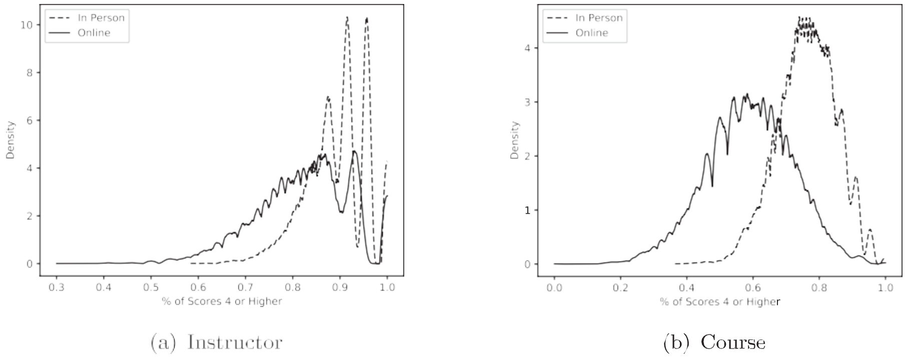
In this section, we analyze the results from Course 5 using the same approach as Course 1. Unlike Course 1, Course 5 experienced a low in-person response rate (32%). Given this difference, the results are expected.

Mean SET scores

Examining Figure 5 shows the width of the distributions by administration method. In this case, it is clear there are thicker tails on the in-person administration. However, all of the distributions are very wide (Figure 6).

While it is true that the in-person administration of the SETs resulted in a larger distribution of the statistic than the online administration, both methods resulted in distributions that are not robust enough for any reasonable use. For instance, at 30 students, the width of the online 95% confidence intervals, for both the instructor and course question, is about three-quarters of a point. This is a large enough difference that, based on current practices at one of the institutions involved in this study, some well-performing adjuncts would not be rehired under the belief they were poor teachers (e.g. an instructor of a 30 person class could receive a '4.00' average score in one semester and a '3.25' in another while holding instruction constant).

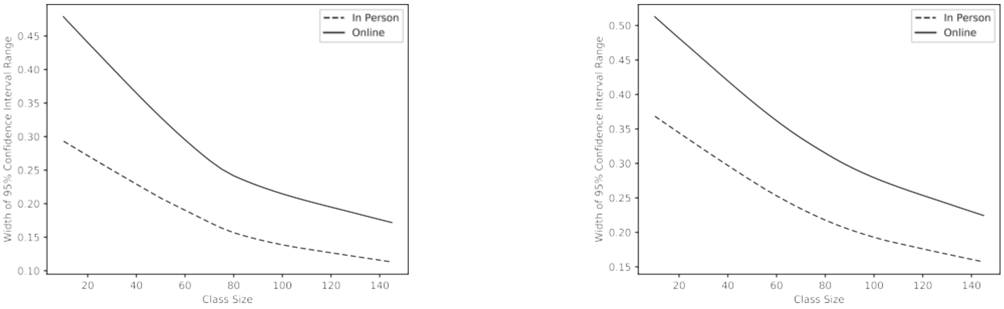
Figure 3. Course 1: KDE of the proportion of responses at or above 4 in simulated classes of 30 students. KS test *p*-values = 0.000.



(a) Instructor

(b) Course

Figure 4. Course 1: LOWESS curve of the proportion of responses at or above 4 in simulated classes of varying size. KS test *p*- values = 0.000.



(a) Instructor

(b) Course

Percent of ratings above 4 as a statistic

As we saw with the first course, a percentage of scores above or equal to a certain value resulted in continued uncertainty. We find this in the post-COVID vaccine period as well. Examining Figure 7, we can see that the distributions, regardless of administration method, span about half of the possible values. This result is even more pronounced when examining Figure 8.

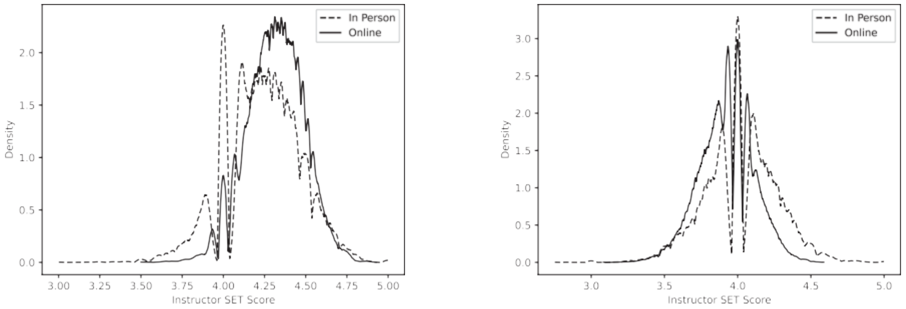
In Figure 8 we see that when examining the narrower of the two distributions, the 95% confidence interval of the statistic for a class of 30 spans at least a third of the available range. Therefore, very little meaning can be derived from this statistic regardless of administration method.

Discussion and conclusion

Student Evaluations of Teaching are frequently used as methods of evaluating the outcomes in a given course, and are commonly used as a means of evaluating the quality of instruction provided in a course. In recent years, the trend has been to move SETs online in an effort to prevent the loss of valuable in-class time that could be spent on increasing learning outcomes. Moreover, COVID has changed the class attendance calculation for many students. These measures are inherently noisy. The amount of noise that can be incorporated into these evaluations at fairly typical class sizes (between 30 and 50 students) is such that the true measure of outcomes (for either instructors or courses) may be impossible to discern based on the ratings provided by students.

The statistical method employed in this study guarantees that all students received exactly the same teaching; all of the students were, in fact, in the same classroom with the same instructor at the same time. Therefore, our results cannot be driven by perceived differences in teaching quality. Moreover, the classes are large enough that the class sample (in contrast to the simulated sample) closely approximates the population distribution for each of the methods of administration. However, this method has one limitation: we assume that the difference in response rate by method of administration does not change with class size. If the response rate difference narrowed in smaller classes, it is possible the difference in the 95% confidence interval by method of administration would decrease. Nonetheless, the data from previous studies (such as Dommeyer et al., 2004) do not indicate that a difference in response rate would be substantial. In our current era where response

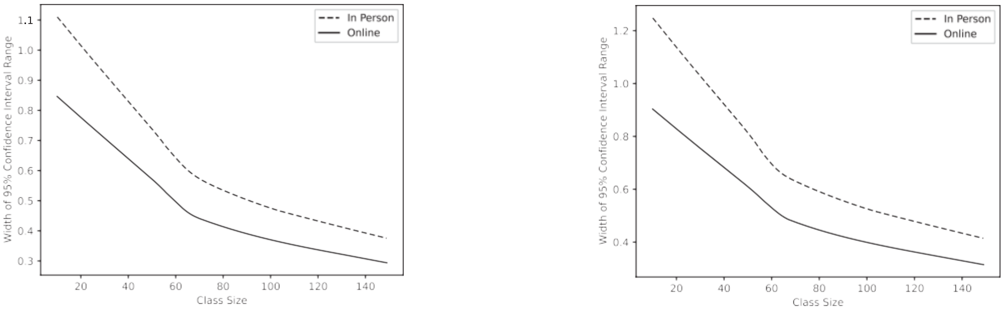
Figure 5. Course 5: KDE for mean SET scores in simulated classes of 30 students. KS test *p*-values = 0.000.



(a) Instructor

(b) Course

Figure 6. Course 5: LOWESS curve for mean SET score in simulated classes of varying size. KS test *p*-values = 0.000.



(a) Instructor

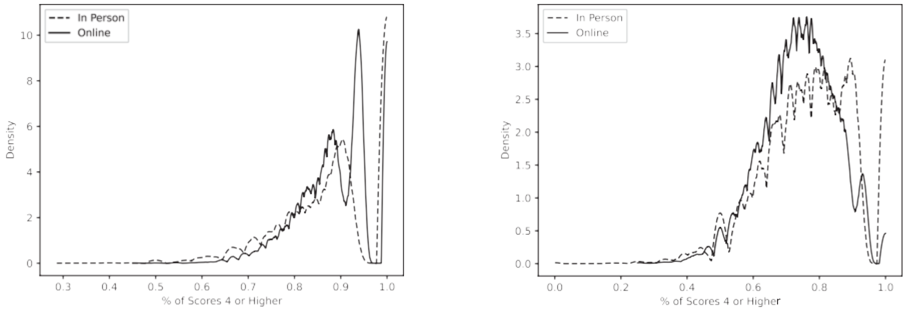
(b) Course

rates are low regardless of administration method, whether there is a difference in the distribution of the statistic based on survey administration method is less pressing. Instead we see a pattern where the true statistic values are uncertain regardless of administration method in the post-COVID vaccine period.

Policy implications

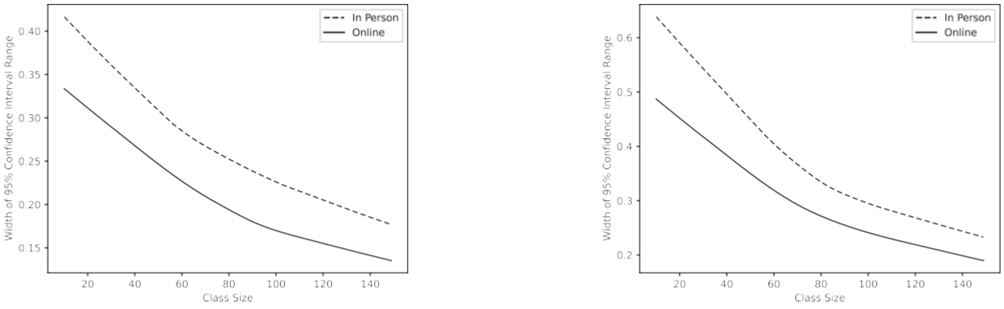
To this point in the paper, we have focused on the statistical difficulties in using SET scores to evaluate outcomes. Given the width of the distribution of the statistic (mean or proportion), it is nearly impossible to use these measures with any helpful degree of accuracy regardless of administration method. This effect is particularly pronounced in smaller classes where the width of the distribution of the statistic (mean or proportion) is larger.

Figure 7. Course 5: KDE of the proportion of responses at or above 4 in simulated classes of 30 students. KS test *p*-values = 0.000.



(a) Instructor (b) Course

Figure 8. Course 5: LOWESS curve of the proportion of responses at or above 4 in simulated classes of varying size. KS test *p*- values = 0.000 for all course values.



(a) Instructor (b) Course

SET scores also face many other problems that reduce their usefulness in accurately evaluating learning outcomes. Boring (2017), MacNell, Driscoll, and Hunt (2015), Arbuckle and Williams (2003), and Basow (1995) find that women are discriminated against in SET scores, especially by their male students. Chisadza, Nicholls, and Yitbarek (2019) and Fan et al. (2019) make similar findings, but find that both race *and* gender are associated with biased scores. Moreover, a substantial number of studies have found little correlation or even an inverse correlation between SET scores and learning (Clayson, 2009, Carrell and West, 2010, Braga, Paccagnella, and Pellizzari, 2014).

The use of SET scores as external metrics of classroom outcomes is deeply entrenched at many universities, and researchers have struggled to disabuse administrators of the usefulness of SET scores as performance measures of instructors. The title of one aptly-named manuscript illuminates the problem of using SET scores: 'Significant interpretation of small mean differences in student evaluations of teaching despite explicit warning to avoid overinterpretation' (Boysen, 2015). Our experiment shows that mean or proportion SET scores are subject to high levels of volatility. In courses with small to moderate numbers of students, this volatility makes almost any interpretation of these scores an overinterpretation.

Note

1. Throughout this paper, we will show KDEs of classes of 30 students. We selected 30 student classes as we believe many administrators believe such a class is sufficient in size to trust the SET scores. This is the default class size of many MBA and upper-division courses at one of the institutions involved in the study.

Disclosure statement

The authors report there are no competing interests to declare.

References

Arbuckle, Julianne, and Benne D Williams. 2003. "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations." *Sex Roles*

49 (9): 507–516. doi:10.1023/A:1025832707002.

Basow, Susan A. 1995. “Student Evaluations of College Professors: When Gender Matters.” *Journal of Educational Psychology* 87 (4): 656–665. doi:10.1037/0022-0663.87.4.656.

Becker, William E., William Bosshardt, and Michael Watts. 2012. “How Departments of Economics Evaluate Teaching.” *The Journal of Economic Education* 43 (3): 325–333. doi:10.1080/00220485.2012.686826.

Becker, William E., and Michael Watts. 1999. “How Departments of Economics Evaluate Teaching.” *American Economic Review* 89 (2): 344–349. doi:10.1257/aer.89.2.344.

Boring, Anne. 2017. “Gender Biases in Student Evaluations of Teaching.” *Journal of Public Economics*145: 27–41. doi:10.1016/j.jpubeco.2016.11.006.

Boysen, Guy A. 2015. “Significant Interpretation of Small Mean Differences in Student Evaluations of Teaching Despite Explicit Warning to Avoid Overinterpretation.” *Scholarship of Teaching and Learning in Psychology* 1 (2): 150–162. doi:10.1037/stl0000017.

Braga, Michela, Marco Paccagnella, and Michele Pellizzari. 2014. “Evaluating Students’ Evaluations of Professors.” *Economics of Education Review* 41: 71–88. doi:10.1016/j.econedurev.2014.04.002.

Breitbach, Elizabeth, Chandini Sankaran, and Jamie Wagner. 2016. “The Movement to Online Course Evaluations: Do We Hear From More Complainers.” *Perspectives on Economic Education Research* 10 (1): 41–56.

Carrell, Scott E., and James E West. 2010. “Does Professor Quality Matter? Evidence From Random Assignment of Students to Professors.” *Journal of Political Economy* 118 (3): 409–432. doi:10.1086/653808.

Chisadza, Carolyn, Nicky Nicholls, and Eleni Yitbarek. 2019. “Race and Gender Biases in Student Evaluations of Teachers.” *Economics Letters* 179: 66–71. doi:10.1016/j.econlet.2019.03.022.

Clayson, Dennis E. 2009. “Student Evaluations of Teaching: Are they Related to what Students Learn? A Meta-analysis and Review of the Literature.” *Journal of Marketing Education* 31 (1): 16–30. doi:10.1177/0273475308324086.

Cleveland, William S., and Susan J Devlin. 1988. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *Journal of the American Statistical Association* 83 (403): 596–610. doi:10.1080/01621459.1988.10478639.

Conover, W. J.. 1998. *Practical Nonparametric Statistics*. 3rd ed. New York: Wiley. Dommeyer, Curt J., Paul Baum, Kenneth S. Chapman, and Robert W Hanna. 2002. "Attitudes of Business Faculty Towards Two Methods of Collecting Teaching Evaluations: Paper Vs. Online." *Assessment & Evaluation in Higher Education* 27 (5): 455–462. doi:10.1080/0260293022000009320.

Dommeyer, Curt J, Paul Baum, Robert W. Hanna, and Kenneth S Chapman. 2004. "Gathering Faculty Teaching Evaluations by in-class and Online Surveys: Their Effects on Response Rates and Evaluations." *Assessment & Evaluation in Higher Education* 29 (5): 611–623. doi:10.1080/02602930410001689171.

Donovan, Judy, Cynthia E. Mader, and John Shinsky. 2010. "Constructive Student Feedback: Online Vs. Traditional Course Evaluations." *Journal of Interactive Online Learning* 9 (3): 283–296. https://www.learntechlib.org/p/109396/.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press.

Fan, Y., L. J. Shepherd, E. Slavich, D. Waters, M. Stone, R. Abel, and E. L Johnston. 2019. "Gender and Cultural Bias in Student Evaluations: Why Representation Matters." *PLoS ONE* 14 (2): Article e0209749. doi:10.1371/journal.pone.0209749.

Fike, David S., Denise J. Doyle, and Robert J Connelly. 2010. "Online Vs. Paper Evaluations of Faculty: When Less Is Just As Good." *Journal of Effective Teaching* 10 (2): 42–54. https://eric.ed.gov/?id=EJ1092118.

Gamliel, Eyal, and Liema Davidovitz. 2005. "Online Versus Traditional Teaching Evaluation: Mode Can Matter." *Assessment & Evaluation in Higher Education* 30 (6): 581–592. doi:10.1080/02602930500260647.

Guder, Faruk, and Mary Malliaris. 2010. "Online and Paper Course Evaluations." *American Journal of Business Education (AJBE)* 3 (2): 131–138.

doi:10.19030/ajbe.v3i2.392.

He, Jun, and Lee A Freeman. 2021. “Can We Trust Teaching Evaluations when Response Rates are Not High? Implications From a Monte Carlo Simulation.” *Studies in Higher Education* 46 (9): 1934–1948. doi:10.1080/03075079.2019.1711046.

Heath, Nicole M., Steven R. Lawyer, and Erin B Rasmussen. 2007. “Web-based Versus Paper-and-pencil Course Evaluations.” *Teaching of Psychology* 34 (4): 259–261. doi:10.1080/00986280701700433.

MacNell, Lillian, Adam Driscoll, and Andrea N Hunt. 2015. “What’s in a Name: Exposing Gender Bias in Student Ratings of Teaching.” *Innovative Higher Education* 40 (4): 291–303. doi:10.1007/s10755-014-9313-4.

McCullough, B. D., and Darrell Radson. 2011. “Analysing Student Evaluations of Teaching: Comparing Means and Proportions.” *Evaluation & Research in Education* 24 (3): 183–202. doi:10.1080/09500790.2011.603411.

Morrison, Keith. 2013. “Online and Paper Evaluations of Courses: a Literature Review and Case Study.” *Educational Research and Evaluation* 19 (7): 585–604. doi:10.1080/13803611.2013.834608.

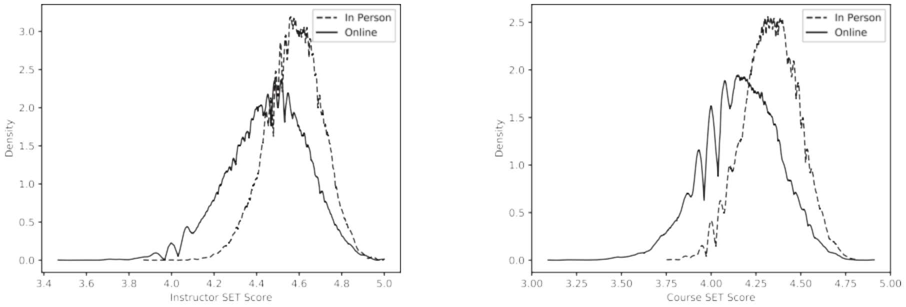
Sankaran, Chandini, Jamie Wagner, and Elizabeth Breitbach. 2018. “Rethinking How We Evaluate Teaching Effectiveness: Does Only the Mean Matter?.” *Perspectives on Economic Education Research*11 (1): 73–87.

Scott, David W. 2015. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, NJ: John Wiley & Sons.

Appendix. Additional simulation results

Course 2

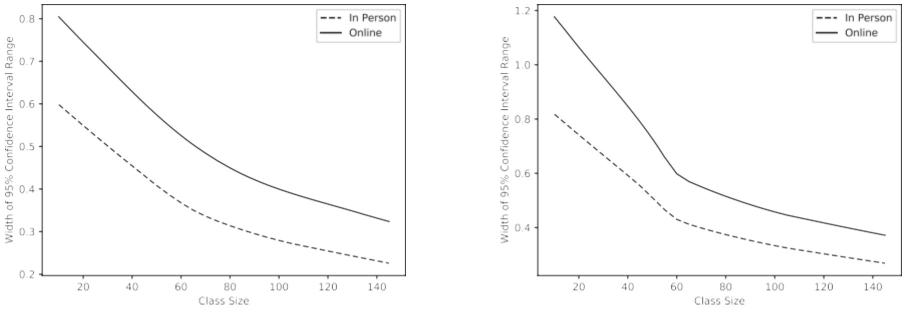
Mean SET scores



(a) Instructor

(b) Course

Figure A1. KDE for mean SET scores in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

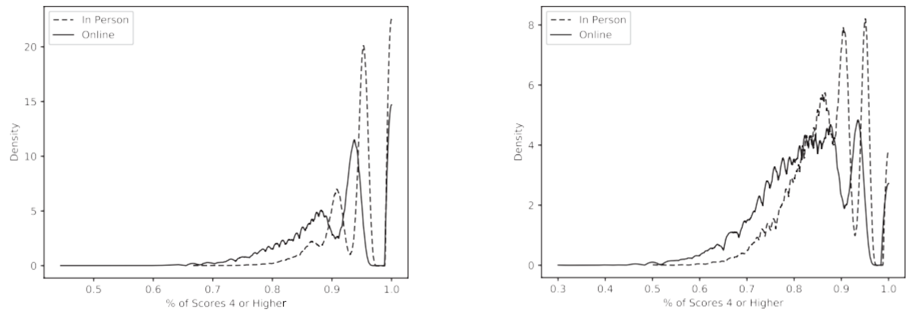


(a) Instructor

(b) Course

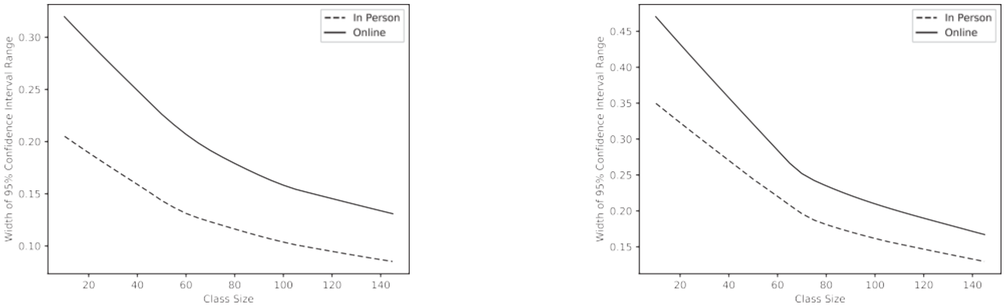
Figure A2. LOWESS curve for mean SET score in simulated classes of varying size. KS test *p*-values= 0.000. (a) Instructor. (b) Course.

Percent of ratings above 4 as a statistic



(a) Instructor (b) Course

Figure A3. KDE of the proportion of responses at or above 4 in simulated classes of 30 Students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

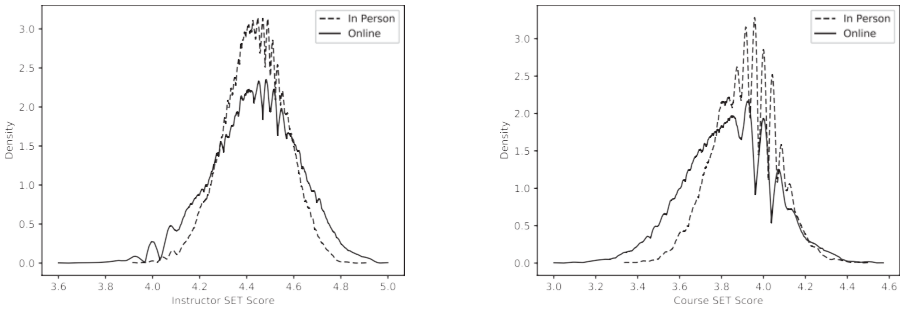


(a) Instructor (b) Course

Figure A4. LOWESS curve of the proportion of responses at or above 4 in simulated classes of varying size. KS test *p*-values
= 0.000. (a) Instructor. (b) Course.

Course 3

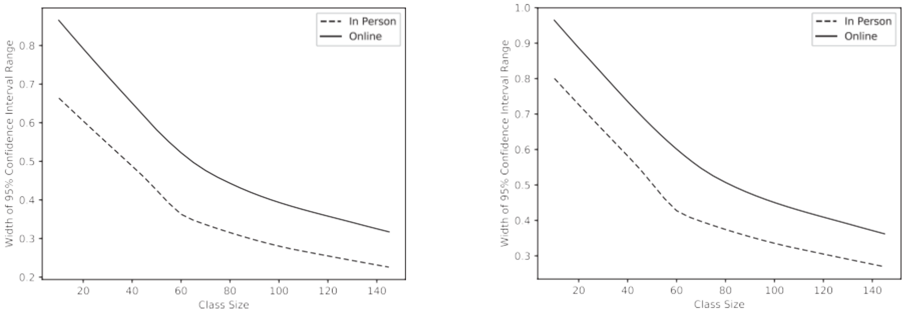
Mean SET scores



(a) Instructor

(b) Course

Figure A5. KDE for mean SET scores in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

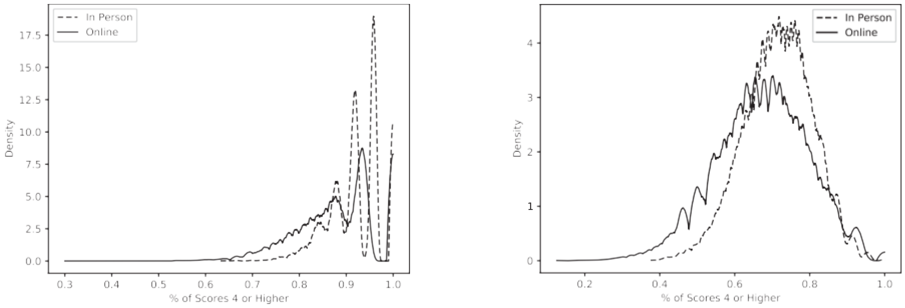


(a) Instructor

(b) Course

Figure A6. LOWESS curve for mean SET score in simulated classes of varying size. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

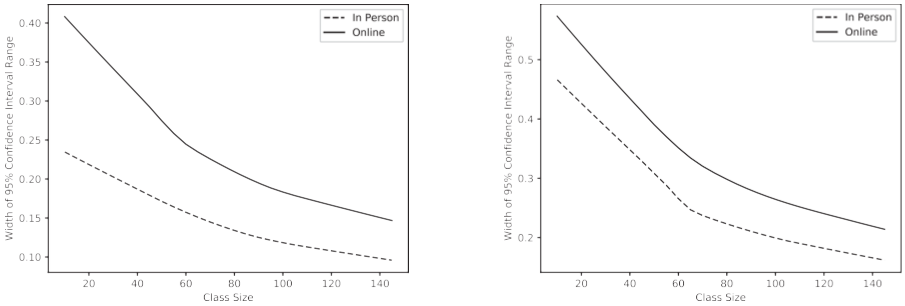
Percent of ratings above 4 as a statistic



(a) Instructor

(b) Course

Figure A7. KDE of the proportion of responses at or above 4 in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.



(a) Instructor

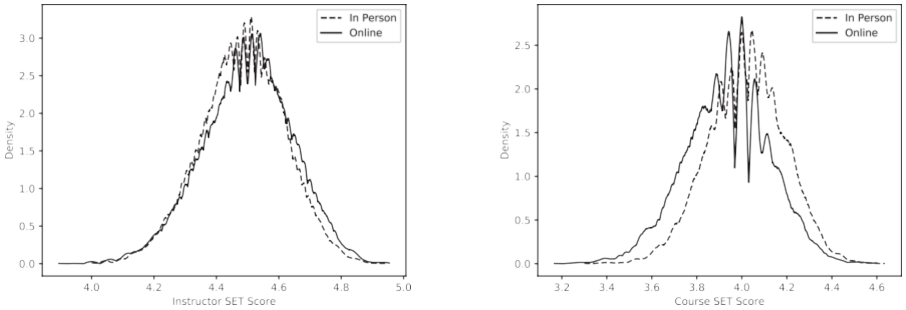
(b) Course

Figure A8. LOWESS curve of the proportion of responses at or above 4 in simulated classes of varying size. KS test *p*-values

= 0.000. (a) Instructor. (b) Course.

Course 4

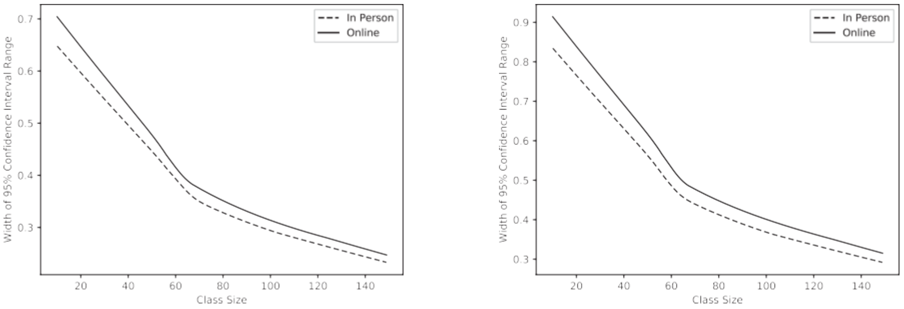
Mean SET scores



(a) Instructor

(b) Course

Figure A9. KDE for mean SET scores in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

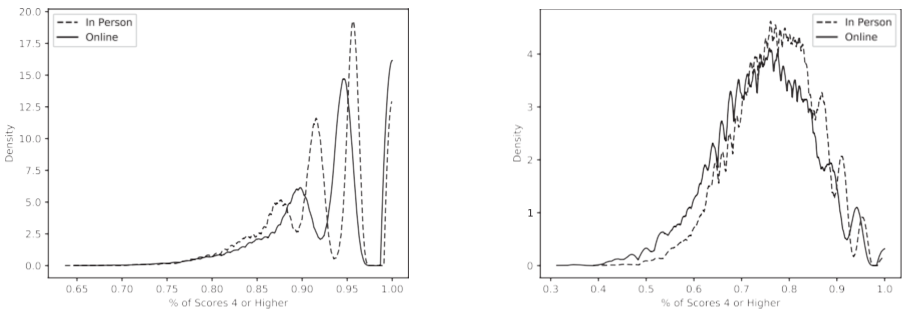


(a) Instructor

(b) Course

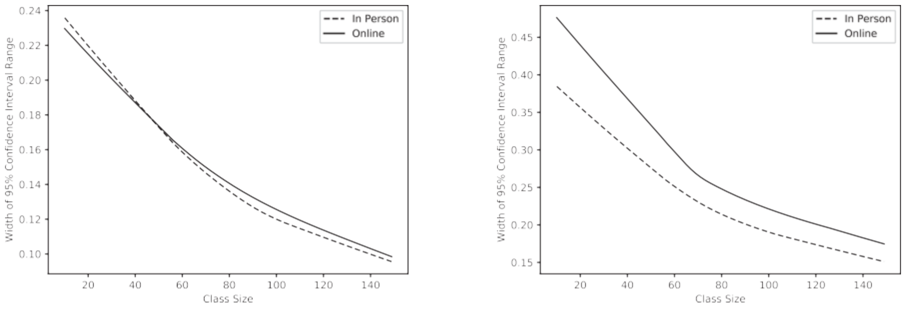
Figure A10. LOWESS curve for mean SET score in simulated classes of varying size. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

Percent of ratings above 4 as a statistic



(a) Instructor (b) Course

Figure A11. KDE of the proportion of responses at or above 4 in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

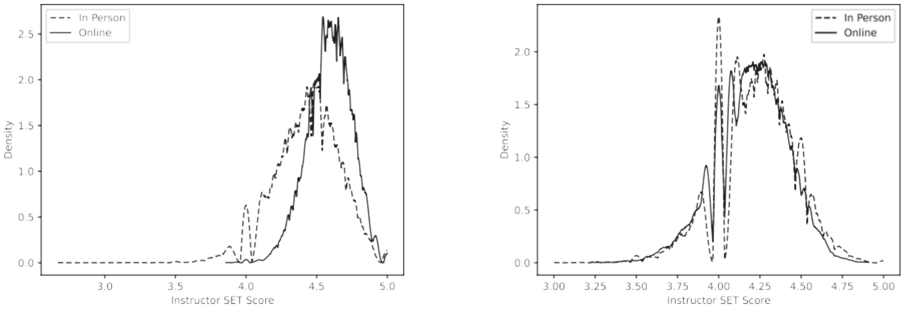


(a) Instructor (b) Course

Figure A12. LOWESS curve of the proportion of responses at or above 4 in simulated classes of varying size. KS test *p*-values = 0.000 for all course values. (a) Instructor. (b) Course.

Course 6

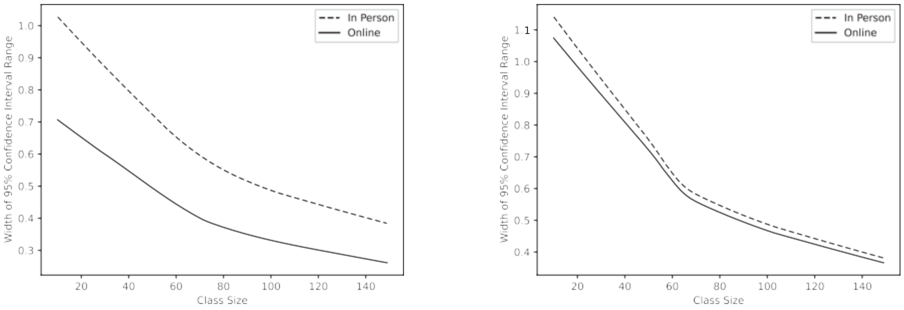
Mean SET scores



(a) Instructor

(b) Course

Figure A13. KDE for mean SET scores in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

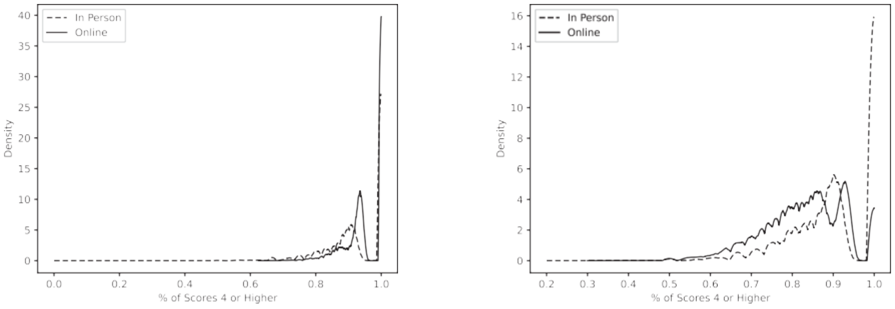


(a) Instructor

(b) Course

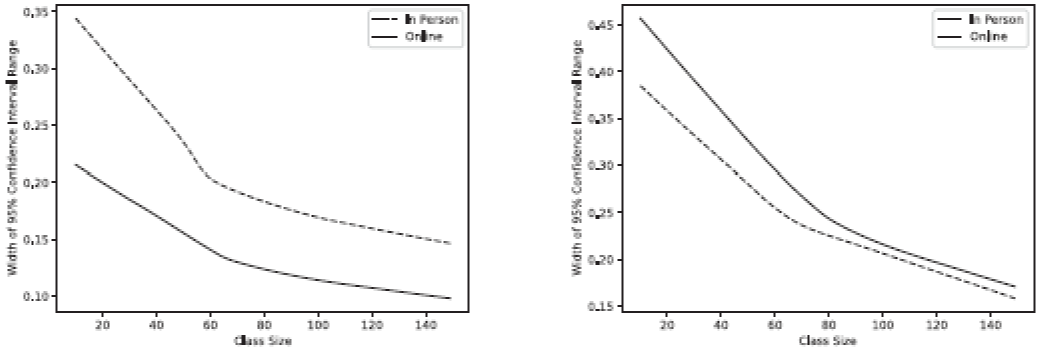
Figure A14. LOWESS curve for mean SET score in simulated classes of varying size. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

Percent of ratings above 4 as a statistic



(a) Instructor (b) Course

Figure A15. KDE of the proportion of responses at or above 4 in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

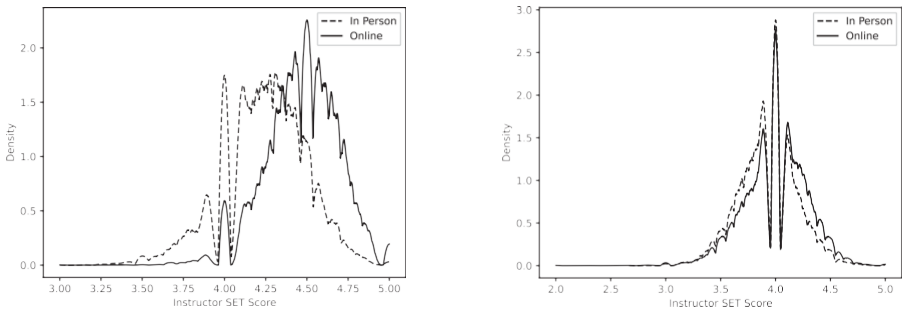


(a) Instructor (b) Course

Figure A16. LOWESS curve of the proportion of responses at or above 4 in simulated classes of varying size. KS test *p*-values= 0.000 for all course values. (a) Instructor. (b) Course.

Course 7

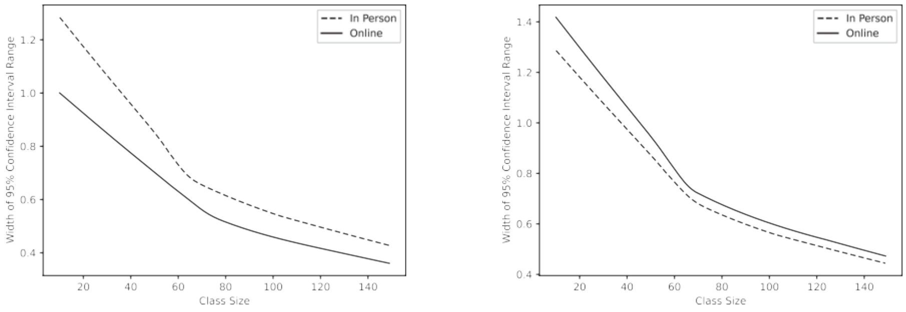
Mean SET scores



(a) Instructor

(b) Course

Figure A17. KDE for mean SET scores in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.

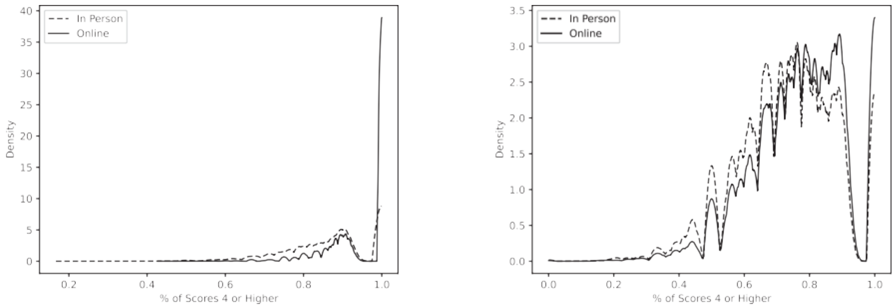


(a) Instructor

(b) Course

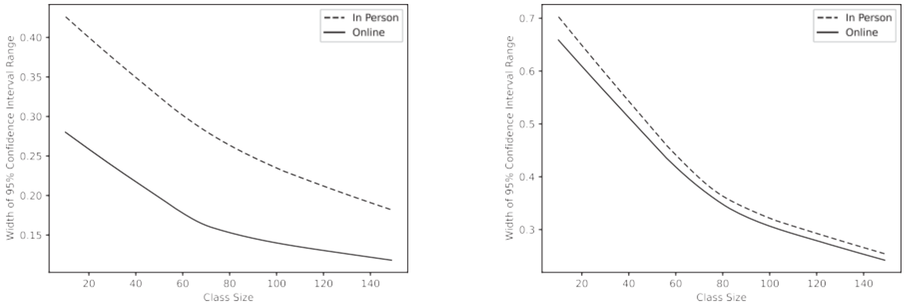
Figure A18. LOWESS curve for mean SET score in simulated classes of varying size. KS test *p*-values= 0.000. (a) Instructor. (b) Course.

Percent of ratings above 4 as a statistic



(a) Instructor (b) Course

Figure A19. KDE of the proportion of responses at or above 4 in simulated classes of 30 students. KS test *p*-values = 0.000. (a) Instructor. (b) Course.



(a) Instructor (b) Course

Figure A20. LOWESS curve of the proportion of responses at or above 4 in simulated classes of varying size. KS test *p*- values= 0.000 for all course values. (a) Instructor. (b) Course.