Unbiased, Reliable, and Valid Student Evaluations Can Still Be Unfair*

Forthcoming in *Assessment & Evaluation in Higher Education*

Justin Esarey[†] and Natalie Valdes[‡]

January 23, 2020

Abstract

Scholarly debate about Student Evaluations of Teaching (SETs) often focuses on whether SETs are valid, reliable, and unbiased. In this paper, we assume the most optimistic conditions for SETs that are supported by the empirical literature. Specifically, we assume that SETs are moderately correlated with teaching quality (student learning and instructional best practices), highly reliable, and do not systematically discriminate on any instructionally irrelevant basis. We use computational simulation to show that, under ideal circumstances, even careful and judicious use of SETs to assess faculty can produce an unacceptably high error rate: (a) a large difference in SET scores fails to reliably identify the best teacher in a pairwise comparison, and (b) more than a quarter of faculty with evaluations at or below the 20th percentile are above the median in instructional quality. These problems are attributable to imprecision in the relationship between SETs and instructor quality that exists even when they are moderately correlated. Our simulation indicates that evaluating instruction using multiple imperfect measures, including but not limited to SETs, can produce a fairer and more useful result compared to using SETs alone.

Keywords: student evaluations of teaching, tenure and promotion, teaching assessment

*We gratefully acknowledge financial support for this project provided by the Wake Forest University Undergraduate Research and Creative Activities Center (URECA), the WFU Associate Provost for Research, and the WFU Department of Politics and International Affairs. We thank Elizabeth Barre, Joshua Eyler, Bethany Morrison, Fred Oswald, and Arthur Spirling for helpful suggestions and comments related to this project.

†Wake Forest University, Department of Politics and International Affairs, Winston-Salem NC 27109. Corresponding author: justin@justinesarey.com.

‡Wake Forest University, Department of Politics and International Affairs, Winston-Salem NC 27109. E-mail: valdnm16@wfu.edu.

Introduction

Should student evaluations be used to assess how well faculty in higher education are teaching? When scholars debate the utility of student evaluations of teaching (SETs), the discussion typically centres on whether SETs measure an instructor's quality in a valid, reliable, and unbiased way. For example, many studies have examined whether SETs are biased against female instructors (Langbein, 1994; Andersen and Miller, 1997; Martin, 2016; Mitchell and Martin, 2018; Mengel et al., 2018). The literature is both voluminous and discordant, perhaps because there are obvious problems with using biased or invalid SETs to make personnel decisions. But what if SETs are free from these problems? In that case, are SET scores a sound basis for choosing which job candidate to hire or whether a faculty member should be granted tenure?

In this article, our computational simulation shows that using SETs to identify poor teachers can result in an unacceptably high error rate *even under the most optimistic scenarios supported by empirical research*. That is, even if it is correct that SETs are (a) moderately correlated with student learning and/or instructional best practices, (b) reliable, and (c) unbiased, common ways that SETs are used to evaluate faculty teaching performance are unfair under reasonable assumptions about the distribution between SETs and instructor quality. This occurs because there is considerable imprecision in the relationship between SET scores and instructor quality even when there is substantial correlation between the two. This imprecision can come from essentially random and idiosyncratic influences on SET score (such as personality or appearance), or it can come from systematic influences on those scores that are not related to instruction (e.g., bias against faculty members of a certain gender or race). But even when there are no systematic biases, the noise created by idiosyncratic variation in SET score interferes with our ability to use SETs to make correct judgements about a faculty member's teaching.

We use computational simulation because it allows us to examine what happens when SET scores are mapped into administrative judgements about faculty teaching under ideal conditions.

Our approach is similar to venerable theoretical models of screening and assessment from industrial psychology (Taylor and Russell, 1939; Naylor and Shine, 1965; Cascio, 1980; Owen and Li, 1980). To avoid our results being overly dependent on distributional assumptions, we use normal copulas[1] that simulate correlated percentile rankings instead of raw scores (Hofert, 2018). Percentile rankings are always uniformly distributed regardless of the distribution of raw SET scores and are therefore a better choice for modelling many universities with different student evaluation instruments and scoring scales.[2] We simulate SET scores and faculty quality percentiles with varying correlation, then use the simulated scores in several assessment procedures. Specifically, we examine:

1. pairwise comparisons of faculty via SET scores. This mirrors the comparison of job candidates on the basis of their teaching performance or the comparison of a faculty member up for tenure to the teaching record of a recent (un)successful case.

2. comparison of an individual professor's SET scores to the overall population of SET scores from all faculty members. This mirrors a procedure where faculty members who are underperforming relative to their peers (e.g., whose scores are below a certain percentile ranking) are identified for administrative action as part of a tenure case or other systematic review.

Even when the correlation between SET scores and faculty instructional quality is $\rho \approx 0.4$, roughly the largest value supported by empirical literature,[3] a large difference in SET scores (even as much as 30 percentile points) does not reliably identify the best teacher in a pairwise comparison of simulated faculty members. Moreover, over one quarter of faculty with SET scores at or below the

[1]A *copula* is a function that connects individual marginal distribution functions to a joint distribution function, and more specifically is 'a multivariate df [distribution function] with standard uniform univariate margins' (Hofert, 2018, pp. 5-6).

[2]However, specific marginal distributions and a correlation coefficient do not uniquely determine a joint distribution of SET scores and instructional quality; our conclusions are therefore still bound to some distributional assumptions that we believe are reasonably (though not universally) generalizable. This point is discussed further in the Methodology section and in our Conclusion, where we discuss the consequences for practical employment of SET scores by faculty and administrators.

[3]For example, the meta-analysis of Cohen (1981) finds an average correlation of 0.43 between instructor's overall SET score and student learning.

20th percentile are actually better at teaching than the median faculty member in our simulation. Even those with exceptionally high SET scores can be poor teachers: nearly 19% of those with SET scores above the 95th percentile are no better than the median professor at teaching. These findings are confirmed when we repeat our analysis using a bivariate normal distribution to simulate SET scores and faculty quality. The implication of our analysis is that making fair, accurate personnel decisions based on faculty instruction requires a measure of teaching performance that is substantially more related to student learning or instructional best practices than SET scores alone.

Based on our results, we make three recommendations concerning how SETs should be used within universities. First, we advise removing any systematic variance in SET scores explained by non-instructional factors (i.e., biases) via regression adjustment or matched subsample analysis before using these scores for any purpose (as illustrated by Nargundkar and Shrikhande, 2014; see also Benton and Li, 2017, pp. 3-4). This adjustment increases the correlation between evaluation scores and teaching quality, in essence filtering out a source of imprecision in this relationship and therefore reducing the chance of an unfair decision according to our simulations. However, this procedure cannot remove the noise created by idiosyncratic influences on SET scores. Thus, we also believe that a combination of independent evaluators, interviews with students, teaching observations by experts, peer review of instructional materials, and SET scores can give a much more accurate picture of a faculty member's teaching proficiency when SET scores alone would be misleading. Importantly, this is true even when each of these individual measures is noisy or flawed. We show that averaging these multiple forms of evaluation can allow idiosyncratic variation in each one to cancel out, resulting in further reduction of imprecision between the averaged assessment and a faculty member's true teaching performance. Because obtaining multiple measures of instructional quality is expensive, we think that SETs could serve as a low-cost mechanism for identifying faculty in need of this more comprehensive assessment of their teaching—with the understanding that many faculty so identified will be excellent teachers. Finally, we advise caution in over-reliance on SET scores for any purpose: we show that their usefulness can depend on characteristics of the joint distribution between SETs and instructor quality that will typically be unknown to administrators or faculty. We believe this final point further underscores the need to use multiple assessment mechanisms (including but not limited to SET scores) when evaluating teaching.

Background

Our study enters a literature that is both large and divided in its assessment of the value of SETs, but essentially united in that its focus is on the validity, reliability, and (un)biasedness of SET scores as measures of faculty teaching performance. In favour of SETs, an influential meta-analysis combining the findings of 41 empirical studies argues that 'we can be relatively certain that the general course and instructor dimensions relate quite strongly to student achievement' (Cohen, 1981, p. 298) and that the 'findings can be generalized to different students, instructors, institutions, and subject matter areas' (p. 305). Re-analysis of the same data by Feldman (1989) largely confirms these conclusions. A long record of scholarship by Herbert Marsh (summarized in Marsh, 1984, 1987; Marsh and Roche, 1997) argues that SETs[4] are:

> (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instruction of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases... and (f) useful in improving teaching effectiveness (Marsh and Roche, 1997, p. 1187).

These findings were confirmed in an independent study by Nargundkar and Shrikhande (2012). Along the same line, Benton and Li (2017, pp. 7-8) reports that student ratings of instruction on a SET produced by IDEA 'correlate[s] positively with external student ratings of learner and

[4]These studies mostly (but not exclusively) examine the reliability and validity of the Student Evaluation of Educational Quality (SEEQ) developed by Marsh (1982).

teacher behaviours' as well as student performance on exams (Benton et al., 2013). These ratings also break into multiple dimensions that correspond to student proficiencies (Li et al., 2016) similar to those reported for the SET created by Marsh (1982). Generally, empirical studies report correlations between student learning and SET ratings of around $\rho \approx 0.4$ at most.

On the other hand, many other studies have argued that SETs do not measure student learning or instructor quality and may be systematically biased against some groups irrespective of merit. Uttl et al. (2017) re-analyse the data from Cohen (1981) and Feldman (1989) and find that their results are influenced by the tendency for small studies with exaggeratedly positive results to be published while null results are ignored (Schwarzer et al., 2015, Chapter 5; see also Sterling, 1959; Sterling et al., 1995; Franco et al., 2014). When examining only those SET studies that included 30 or more class sections, the correlation between student ratings and student learning falls by 50% or more (see Table 3 in Uttl et al., 2017); when including SET studies published later, this correlation falls even further to at most $\rho \approx 0.1$ (see Table 4). While Nargundkar and Shrikhande (2014) confirm the validity of Marsh's (1982, p. 55) SET instrument, they also note that SET scores 'reflect some inherent biases due to noninstructional factors' (p. 55) such as class size, instructor gender, and the type of course being taught. Work on SETs in our home discipline (of political science) has been particularly focused on the effect of instructor gender on course evaluations, repeatedly finding that women receive lower SET scores than men even when the other aspects of the course are equivalent (Langbein, 1994; Andersen and Miller, 1997; Martin, 2016; Mitchell and Martin, 2018). This finding was partially confirmed in a large cross-disciplinary study by Mengel et al. (2018), who found bias against female instructors but only against those early in their career.

Using invalid, unreliable, or biased student evaluations to make decisions about hiring and tenure is obviously harmful to students and faculty alike. If universities use SETs that do not track student learning or instructional best practices as a part of their hiring, tenure, and promotion process, they may make important decisions on the basis of random chance instead of merit. Even worse, biased SETs could disadvantage faculty from underrepresented minority groups or punish faculty members who teach unpopular required courses. While we agree that this is an important problem, we ask a different question: if SETs are valid, reliable, and unbiased, what then? Are SET scores without demonstrable bias and moderately correlated with instructor quality a fair basis on which to judge a faculty member's teaching performance? If the answer to this question is 'no,' there is a much bigger problem with the use of SETs than is commonly recognized.

Methodology

Rather than empirically assessing the reliability, validity, or unbiasedness of SET scores as measures of teaching proficiency, we *assume* the most optimistic possible conditions that can be supported by empirical literature and then study the outcomes of using SET scores as tools for assessing and managing university faculty. First, we assume that overall instructor SET ratings are moderately correlated with actual instructional quality (e.g., student learning and/or instructional best practices) in the course. The highest correlations reported in the literature are on the order of $\rho \approx 0.4$, so we use this level of correlation as the basis for our study. We also assume that an instructor's class average SET scores are perfectly reliable (i.e., an instructor's class average SET score is always the same and thus perfectly known). This is equivalent to assuming that a faculty member is assessed using the average of a large number of class-average SET scores, or alternatively that class average SET scores are perfectly stable. Second, we assume that administrators use SET scores in the careful and judicious manner recommended by experts (e.g., Boysen et al., 2014; Benton and Young, 2018). For example, we assume that administrators require a large difference in SET scores before concluding that there is a real difference in quality between two instructors, given the imperfect correlation of SET scores with quality reported by the literature. Relatedly, if instructors who receive lower SET scores than their peers are identified as poor teachers, we assume that administrators set this cutoff to a low level in order to minimize the possibility for error. Finally, although we model the idiosyncratic difference in instructor class-average SET scores that is unrelated to student learning, we assume that these idiosyncratic differences are both stable and not representative of any systematic bias against a particular group (e.g., women or under-represented minorities).

These assumptions inform our model of SET scores, the quality of instruction, and administrative assessment. Specifically, we study the percentile rankings (rather than raw scores or values) of a faculty member i's class average overall instructor SET scores s_i and their actual quality as a teacher q_i, where q_i represents a faculty member's true contribution to student learning and/or their conformity with instructional best practices. For example, $s_i = 40$ means that the faculty member's overall SET score is better than 40% of SET scores from the total population of faculty members, while $q_i = 40$ means that the same faculty member is a better instructor than 40% of his/her peers. Because these measures are percentiles, each has a uniform marginal distribution regardless of how the underlying raw SET score or instructor quality metric is distributed: $s_i \sim U[0, 100]$ and $q_i \sim U[0, 100]$. The *joint* distribution of (s_i, q_i) is such that s_i and q_i are correlated at a fixed and common value of ρ; we model this using a normal copula with correlation ρ (Hofert, 2018, p. 87).

Figure 1 shows three examples of how we simulate faculty members' SET scores and true instructional quality values as percentiles using our model. Each point in a scatter plot (s_i, q_i) represents an individual faculty member i's class average SET score percentile (on the x-axis) and true instructional quality percentile (on the y-axis). In Panel 1a, there is no relationship between SET score and true instructor quality ($\rho = 0$). In Panel 1b, the correlation between SET score and instructor quality is set at a value roughly consistent with the maximum value supported by empirical literature ($\rho = 0.4$). Finally, Panel 1c sets the correlation extremely high ($\rho = 0.9$). The histograms on the top and right edges of each scatter plot shows the marginal uniform distribution of SET score percentiles (top edge) and faculty instructional quality (right edge). Each one of our simulations draws a large number of faculty members from the normal copula with a specified ρ and uses those simulated faculty members to assess how well particular administrative uses of SET scores work in terms of distinguishing good teachers from bad. Because we know the true instructor quality of each faculty member in our simulations, we are able to accurately assess how well these administrative uses perform under ideal conditions.

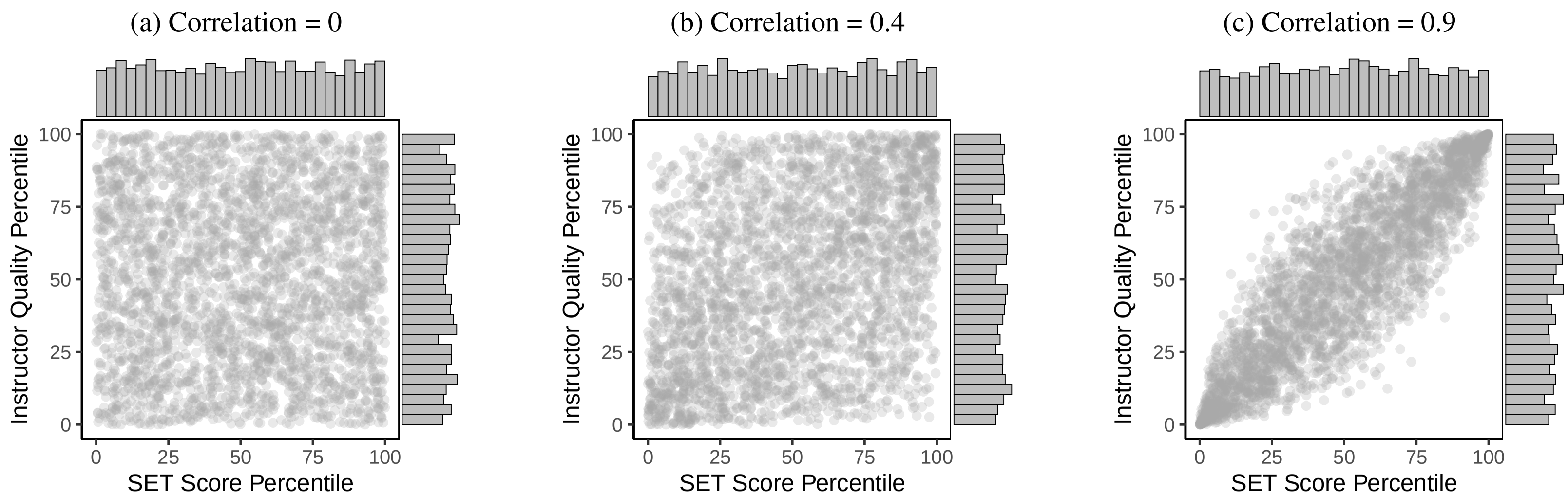
Our model is similar to that of Taylor and Russell (1939), who studied the consequences of using a standardized assessment (such as a scored test) to select workers when this assessment is meaningfully but imperfectly correlated (at level r) with those workers' job performance. Because Taylor and Russell were working at a time before low-cost electronic computing, their main contribution was producing tables showing what proportion of workers selected by the assessment would be acceptable given the stringency of the assessment (i.e. what proportion of assessed workers would be hired), the validity of the assessment (r), and the proportion of workers in the population whose job performance would be at least minimally acceptable; these tables enabled managers to make informed decisions without complex calculation. For the same reason, Taylor and Russell assumed that the continuous standardized assessment metric (e.g., a test score) and the continuous standardized measure of worker quality would be distributed according to a unit bivariate normal density with correlation r. Naylor and Shine (1965) created a version of these tables showing the average standardized job performance of workers selected under varying assessment conditions; their modified approach allowed managers to determine how much selection via standardized assessment would improve the average quality of a workforce relative to the baseline level of the overall population. Naylor and Shine maintained the assumption of a correlated unit bivariate normal distribution between these standardized quality scores. These models have been successfully applied to the study of assessment in higher education in the past; for example, Owen and Li (1980) study the effectiveness of of requiring minimum standardized test score performance as a requirement for college admissions using a modified Taylor-Russell approach.

We use our model to study two common administrative uses of SET scores:

1. for pairwise comparison of faculty members; and

2. for comparison of an individual faculty member's SET score to the overall distribution of

Figure 1: **Simulated SET Scores and Instructor Quality Levels:** the graphs shows the relationship between simulated class averaged SET score percentiles (s_i, shown on the x-axis) and instructor quality percentile ratings (q_i, shown on the y-axis) from a normal copula with correlation $\rho \in \{0, 0.4, 0.9\}$. Each point in the scatter plot (s_i, q_i) represents a simulated faculty member. Figure 1a shows an example with no correlation between SET score and instructor quality. Figure 1b shows an example with moderate correlation ($\rho = 0.4$) between SET score and instructor quality. Figure 1c shows an example with extremely high correlation ($\rho = 0.9$) between SET score and instructor quality. Each figure shows 3,000 simulated faculty members. The histograms on the top and right edges of each scatter plot show the uniform distribution of percentiles for SET scores (on the top edge) and instructor quality (on the right edge). The data are simulated using the `copula` library (Hofert et al., 2017; Kojadinovic and Yan, 2010) in Microsoft R Open 3.5.3 (R Core Team, 2019).



SET scores from all faculty.

These uses are designed to cover realistic scenarios in which SETs may inform decision-making. For example, tenure evaluations often make reference to cases from the recent past; pairwise comparison of a candidate's SET scores to those of a recent tenure case might be used to justify a decision based on teaching performance. Hiring decisions involve comparing a small number of faculty members to one another; a pairwise comparison of SET scores might be used to adjudicate which candidate is a better teacher. Perhaps most likely of all, a faculty member's SET scores might be compared to the larger population of SET scores from all faculty in order to identify those whose teaching performance is markedly worse than their peers. For example, an instructor might be judged according to whether his/her SET scores are below the department or university median score. We expect that such a comparison is almost guaranteed to happen as part of a tenure review or a pre-tenure evaluation.

We model the impact of these procedures by sampling 1,000,000 draws of SET score percentiles and instructor quality percentiles from a normal copula with correlation ρ using the `copula` library (Hofert et al., 2017; Kojadinovic and Yan, 2010) in Microsoft R Open 3.5.3 (R Core Team, 2019). As illustrated in Figure 1, each draw of a SET score and an instructor quality percentile represents a single faculty member. Then, for the first procedure, we compare the first 500,000 SET score draws to the second 500,000 SET score draws to determine the proportion of the time that the faculty member with the higher SET score is also a higher quality instructor. For the second procedure, we compare all 1,000,000 SET score draws to a minimum percentile threshold, separate all faculty members whose SET score is below that threshold from the population, then examine the instructor quality scores of faculty members who are below the SET percentile threshold.

One methodological choice requires particular attention: our choice to simulate percentiles of SET scores and instructor performance using the normal copula rather than raw or standardized scores.[5] This choice comes with an important advantage: it enables us to avoid assuming any

[5] A standardized measure (i.e., a z-score) is one that has been rescaled to have a mean of zero and a standard deviation

particular marginal distribution of SET scores or instructor quality ratings. However, as we will revisit in the conclusion, a correlation ρ is consistent with many possible relationships between SET score and instructor quality and different relationships result in different consequences of using SET scores to evaluate faculty. Focusing on percentiles rather than raw or standardized scores allows us to sidestep the issue of specifying correct marginal distributions and makes our analysis more broadly applicable, as any SET scores or instructional quality metrics must have the same distribution of percentile ratings. That is, no matter how a university asks its students for an overall instructor rating—for instance, using different question wording or different response scales—the percentiles from those ratings *must* be distributed uniformly and therefore our simulations are better able to model the consequences of using any one of these SETs.

However, there are limitations to our procedure. First, specifying marginal distribution functions and a correlation coefficient does *not* imply a unique joint distribution between SET scores and faculty quality (Hofert, 2018, p. 47): the same uniformly distributed SET scores and instructor quality metrics can have different patterns of relationship that all have the same correlation ρ. Consequently, although we believe that our conclusions generalize to a wide variety of common circumstances, they do not necessarily apply to every possible case. Moreover, the empirical literature often studies the correlation between raw or z-score standardized SET scores and direct measures of student achievement like exam scores (e.g., Benton et al., 2013, pp. 380-383), not the correlation between percentile rankings for these measures; we must assume that this difference is not consequential in parts of our analysis. Consequently, we consider our analysis an informative illustration of what can go wrong under reasonable and likely conditions, not a mirror-like simulation of the outcome of using any particular SET system. Toward the end of increasing the robustness of our conclusions, we repeated our analyses using bivariate normally distributed SET scores and faculty quality metrics instead of uniformly distributed but correlated percentiles created via a normal copula; the results, which are reported in an online appendix, are qualitatively

of one.

similar to our main findings. We will return to a discussion of this methodological choice when describing the implications of our study in the conclusion.

Results

Figure 2 shows the result of conducting pairwise comparisons of faculty members using average SET scores. If these SET scores were assigned purely at random, the faculty member who was truly a poorer instructor would be identified by SET scores as the better instructor 50% of the time (that is, $s_i - s_j$ and $q_i - q_j$ have the same sign for a pair of faculty members i and j).[6] Therefore, a 50% error rate is the minimum baseline of evaluation performance.

Figure 2 shows that comparing faculty members' class average SET scores results in an unacceptably high error rate, even when there is moderate correlation ρ between true quality and overall instructor SET evaluation. Based on the empirical literature, $\rho \approx 0.4$ is the highest correlation we can realistically expect. At this level of correlation between quality and SET scores, the poorer-quality instructor in our simulation has a higher average SET score almost 37% of the time. That is, using SET scores that are perfectly reliable and moderately correlated with teaching quality to compare two faculty members can identify the wrong faculty member as a better teacher over one third of the time. Even when $\rho = 0.9$, far higher than possible in real data, the poorer-quality instructor still has a higher average SET score over 14% of the time.

Scholars who believe that SET scores have a role to play in administrative decisions do not endorse taking very small differences in SET score seriously (Boysen et al., 2014; Benton and Young, 2018). Therefore, we repeated our analysis fixing $\rho = 0.4$ and varied the size of the gap in percentile SET scores necessary to conclude that one faculty member is truly a better instructor than another. The result is depicted in Figure 3.

Requiring a minimum distance between SET scores in order to form a judgement about in-

[6]Throughout our analysis, $s_i = s_j$ is treated as a correct prediction if and only if $q_i = q_j$. In addition, when $q_i = q_j$ only $s_i = s_j$ is treated as a correct prediction. These are all edge cases and computationally unlikely to occur.

Figure 2: **Pairwise Comparison Error Rate:** the graph shows the ability of SET overall instructor scores to identify the better teacher in a pairwise comparison. The x-axis shows ρ, the correlation between SET overall instructor score s_i and actual instructor quality q_i as drawn from a normal copula. The y-axis indicates the estimated proportion of the time that the higher quality instructor also has the lower average evaluation (that is, $s_i - s_j$ and $q_i - q_j$ do not have the same sign). Each point in the plot represents an estimate from 500,000 simulated comparisons. Gray dashed lines indicate the simulated error rate when correlation is 0.4 and 0.9, as discussed in the text.

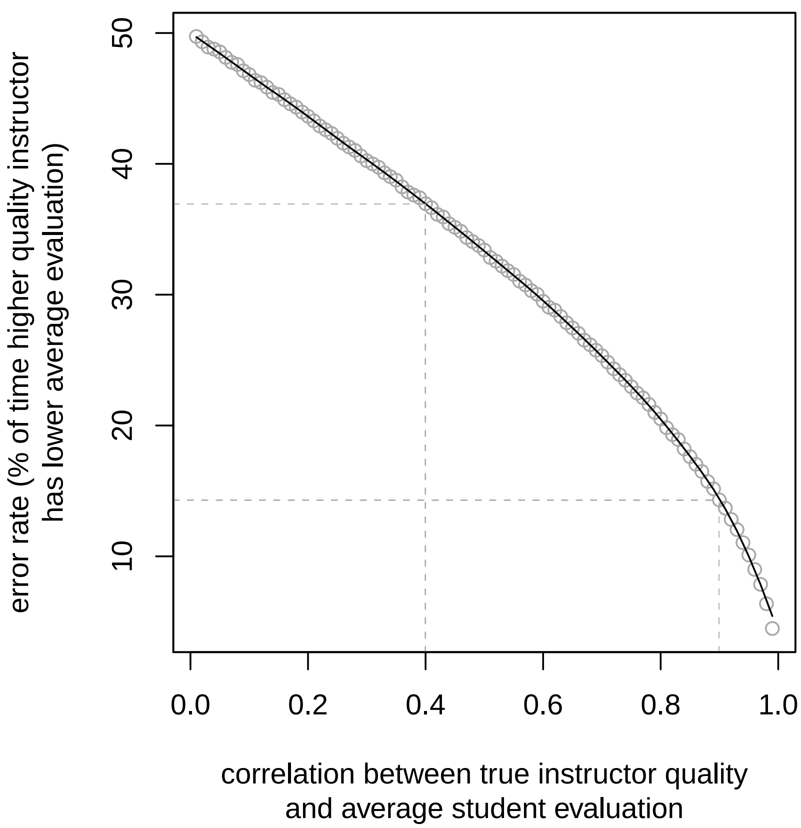
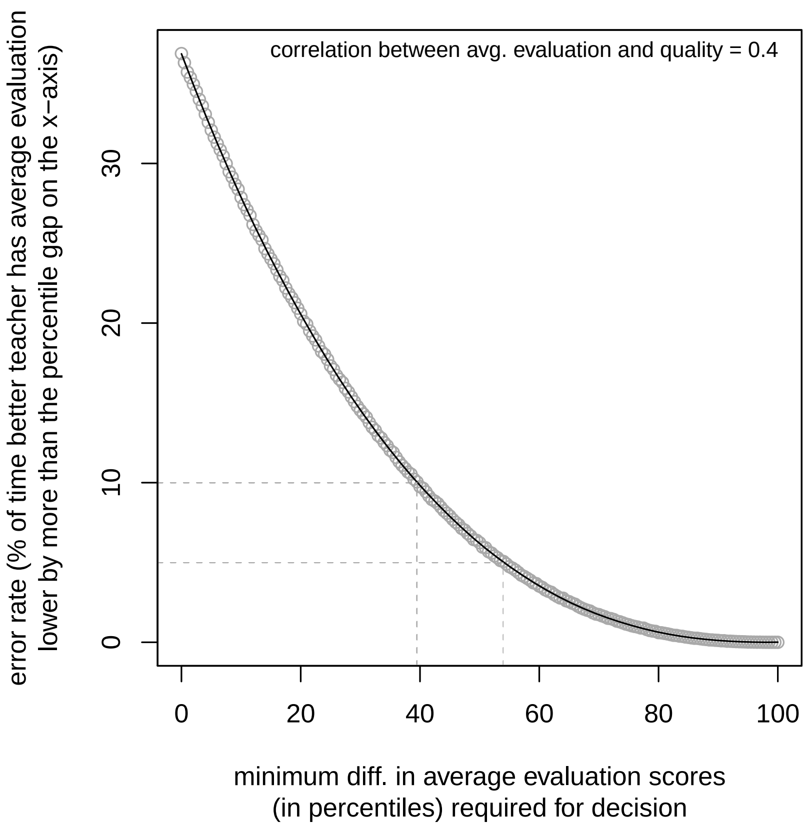


Figure 3: **Pairwise Comparison Error Rate with Minimum Gap:** the graph shows the ability of SET overall instructor scores to identify the better teacher in a pairwise comparison when scores must be at least a certain distance apart in order to conclude that the difference is meaningful. The x-axis shows the minimum distance between SET scores $g_{ij} = (s_i - s_j)$ that must exist before we decide that faculty member i is a better teacher than faculty member j. The y-axis indicates the estimated proportion of the time that the higher quality instructor also has a negative SET score gap larger than the minimum decision value. The correlation between SET overall instructor score and actual instructor quality ρ is fixed at 0.4. Each point in the plot represents an estimate from 500,000 simulated comparisons. Gray dashed lines indcate the approximate minimum difference in evaluation scores needed to achieve 10% and 5% error rates in decision-making.



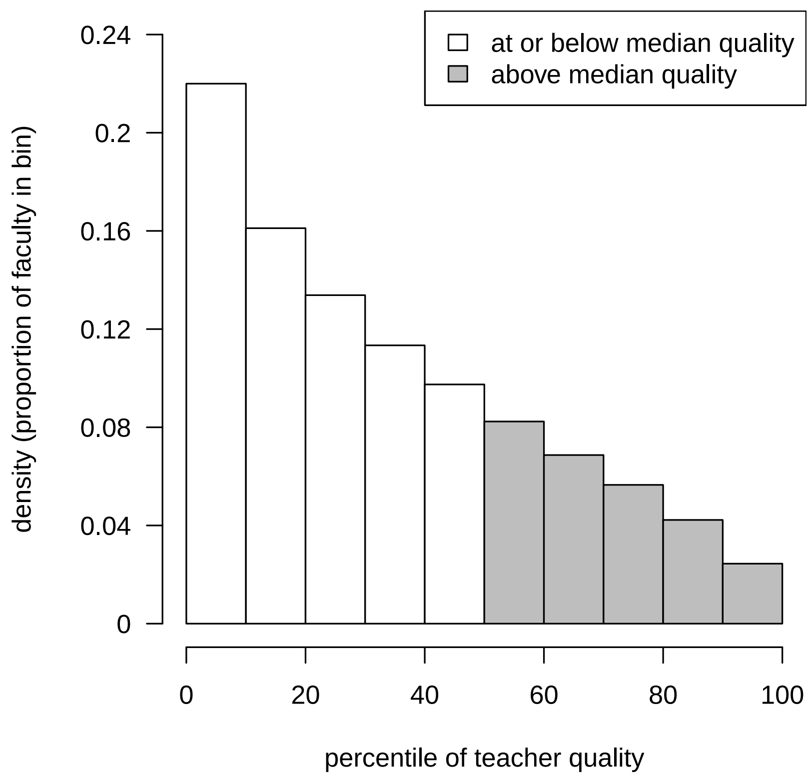
structor quality does improve considerably on simply taking any difference in SET scores seriously. However, the gap in SET scores in our simulation must be very large indeed before errors in decision making reach an acceptable level. As Figure 3 shows, only when our simulated faculty members are separated by a SET score difference of about 40 percentile points does the error rate reach 10%. Even these comparisons identify the wrong instructor as better one out of ten times. A five percent error rate can be achieved in our simulation by comparing only faculty members whose SET scores are separated by about 54 percentile points; however, only about 21% of our simulated pairwise comparisons had a difference in SET scores at least this large, meaning that no decision about which professor was a better teacher could be made in the vast majority of cases.

Finally, faculty may not be explicitly compared to one another but rather compared to the overall population of all faculty as part of a review process; for example, administrators may compare a faculty member's SET scores to the department or university median SET score. We simulated this performance evaluation by drawing 1,000,000 SET scores and true instructor quality levels from a normal copula with correlation fixed at $\rho = 0.4$, then identified faculty members at or below the 20th percentile of SET overall instructor score as low-quality teachers. Using the 20th percentile instead of the median is intended to represent a careful administrative use of SETs that identifies the worst teachers, not simply teachers who are not exceptionally good. Finally, we examine the distribution of *true* instructor quality levels in this population of faculty members with poor SET scores. The result is shown in Figure 4.

Figure 4 shows that, even with moderate correlation between SET scores and true instructor quality, some of the worst performers on student evaluations of teaching are still good teachers in our simulation. Specifically, over 27% of simulated faculty members at or below the 20th percentile on SETs were actually above the median of instructor quality.[7] It is disturbing that even a relatively low floor for SET performance under the most optimistic conditions supported by empirical study

[7]For the distributions we use, the problem is symmetric: that is, over 27% of faculty members with SET scores over the 80th percentile are actually at or below the median in instructor quality.

Figure 4: **Distribution of True Instructor Quality among the Professors with the Poorest SET Scores:** the histogram depicts the distribution of a faculty member's true quality as an instructor (relative to the full population of all faculty members) if that faculty member has a SET overall instructor score at or below the 20th percentile. The bins are spaced such that the height of the bin represents the proportion of faculty members in the bin. Results are based from 1,000,000 draws from the normal copula with correlation between SET score and instructor quality fixed at $\rho = 0.4$. Just over 27% of faculty members at or below the 20th percentile of SET scores are above the median in teacher quality.



results can create a scenario where one in four faculty members identified as a poor teacher is actually more capable than the typical professor.

Indeed, not even exceptional SET scores are a reliable indicator of faculty quality. To simulate the process of selecting the most highly-rated professors for special recognition, we repeated the analysis of Figure 4 and identified simulated faculty members above the 95th percentile in SET scores. We find that over 18% of this elite group of faculty are no better than the median instructor.[8] Based on these results, it would not be unusual to find that a substantial proportion of teaching award winners are actually worse at teaching than the ordinary faculty member. Thus, not only do poor SET scores often mis-identify good professors as bad, they can often identify worse-than-mediocre professors as exceptionally skilled.

Conclusion

Our evidence indicates that common uses for student evaluations of teaching can easily produce many unfair outcomes when those evaluations are extremely reliable, unbiased against any group, and moderately correlated with true faculty quality. As we see it, the fundamental problem is that irrelevant influences on student evaluation scores make decisions based on these scores too subject to chance. At the same time, we recognize that student evaluations have substantial advantages in terms of cost and standardization over available alternatives. What should be done? We make three recommendations.

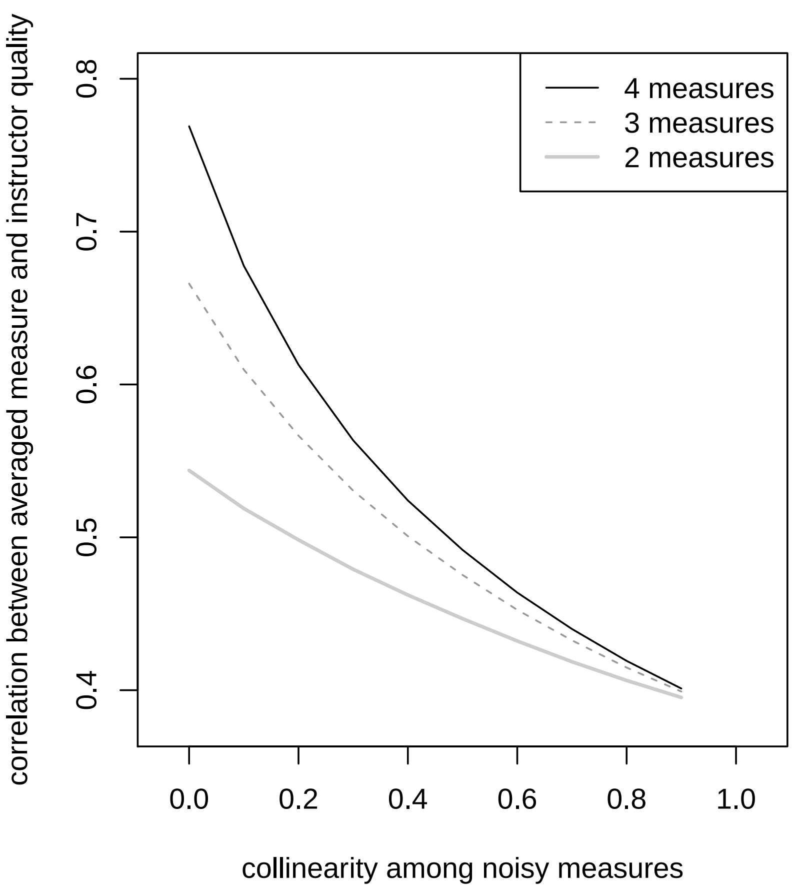
First, consonant with the recommendations already produced by some companies that create SETs (e.g. Benton and Young, 2018) and supported by independent research (e.g., Nargundkar and Shrikhande, 2014), we believe that course-averaged student evaluation scores should be statistically adjusted to remove any systematic non-instructional influences (i.e., biases) before they are

[8] See Appendix Figure 7 for the full distribution of teacher quality ratings for those with extremely high SET scores. As before, the problem is symmetric for the distributions that we examine: over 18% of faculty members with SET scores at or below the 5th percentile are actually above the median in instructor quality.

used for any purpose. For example, linear regression could extract any variance in course average SET scores attributable to gender, race, and course type before these scores are used to evaluate faculty. As long as these characteristics are irrelevant to teaching and learning, the resulting adjusted scores should be more strongly correlated with true instructor quality than the raw SET scores and imprecision will be reduced. As another alternative, scores could be matched on these characteristics and then compared only within the matched groups; this is similar to the approach already taken by IDEA (Benton and Li, 2017, p. 5). This procedure cannot remove purely random or idiosyncratic influences on SET scores, and so even with this adjustment we do not expect that correlation between SET scores and instructor quality will be close to 1. But we believe that the adjustment we recommend will make this correlation as high as it can be.

Second, and also consistent with prior recommendations (Benton and Young, 2018), we believe that SET scores should be used in concert with multiple, dissimilar measures in order to comprehensively evaluate the teaching performance of faculty members when significant personnel decisions are being made. Because every method (including but not limited to SETs) produces a noisy and possibly problematic measure of faculty instructional quality, 'no single measure is adequate as evidence for evaluating teaching' (Benton and Young, 2018, p. 3). However, as long as these measures are not all noisy or biased in the same way, employing them all simultaneously to evaluate teaching can result in a more valid assessment. Consider Figure 5, which shows the correlation between true teacher quality and an average of noisy measures thereof. Each noisy measure is correlated with quality at $\rho = 0.4$, and the measures are correlated with each other as indicated on the x-axis (which we label in the figure as *collinearity*). When collinearity among the measures is low, it indicates that idiosyncratic and instructionally irrelevant influences on these measures are different for each one. When collinearity among the measures is high, it means that the biases and imperfections of each measure are very similar. As the plot shows, the validity of a combined assessment of four different measures can be substantially better than any individual measure as long as these measures are imperfect in different ways.

Figure 5: **Improvement in Measure Quality via Averaging:** the graph shows an estimate of the correlation between the average of multiple measures of instructor quality (on the y-axis) as a function of the degree to which the idiosyncratic, instructionally irrelevant influences on each measure are correlated with one another (labeled *collinearity* on the x-axis). Correlation of the average measure is estimated using 10,000 simulations. In each simulation, a thousand observations of instructor quality and four imperfect measures of quality are drawn from a normal copula. Each measure is correlated with instructor quality at $\rho = 0.4$, and each measure is correlated with the other measures according to a value of ρ (at equally spaced intervals between 0.1 and 0.9) indicated on the x-axis. The measures are averaged and the correlation of the average with true quality is calculated for all 10,000 simulations. The mean correlation across simulations is listed on the plot. The process is repeated using all four measures, only three measures, and only two measures; each is listed on the plot.



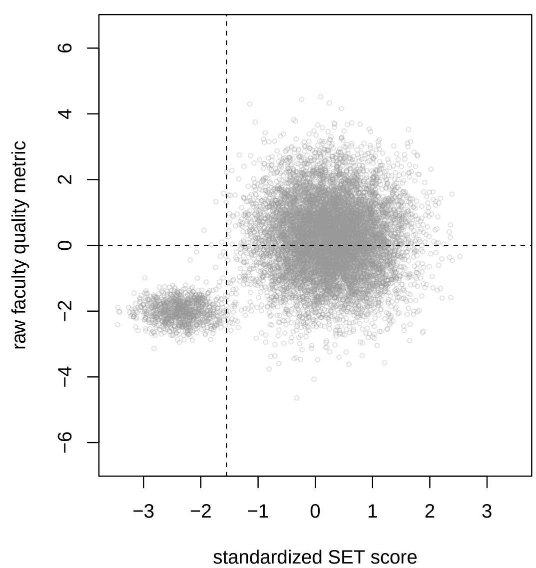
Of course, this kind of comprehensive assessment is very costly; it requires all faculty to engage in self-assessment, peer review of syllabi and other course materials, direct observation of their classroom teaching by trained experts, extensive interviews by administrators of their students, and the like on an annual basis. Therefore, we believe that SET scores could serve as a low-cost means to initially screen some faculty members for this more comprehensive evaluation. Based on our simulation evidence, we think it important to emphasize that many of the faculty members selected for this evaluation will be good teachers (and some faculty members not selected will be poor teachers). Therefore, we believe that *the decision to intensively evaluate a professor's teaching should not be viewed as punitive*. Nor should this intensive evaluation be performed solely on faculty with low SET scores; as we showed in this paper, we expect that some poor instructors will receive good student evaluations by chance and thus we think that a random subset of faculty with good SET scores should be selected for intensive evaluation as well.

Finally, we think that administrators and faculty members should be mindful that the shape of the joint distribution between SET scores and instructor quality can impact the usefulness of SET scores even when the correlation between SET scores and faculty quality is held constant. As noted above, our analysis presumes a correlation between percentile rankings (rather than raw SET scores or faculty quality measures) because percentiles are always uniformly distributed. This makes our analysis applicable to a broader variety of cases. However, the same correlation can correspond to very different distributions. Evaluating a university's faculty according to their performance on SET score percentiles can still be problematic depending on precisely how they are related to teacher quality, which might vary both among SET instruments and universities.

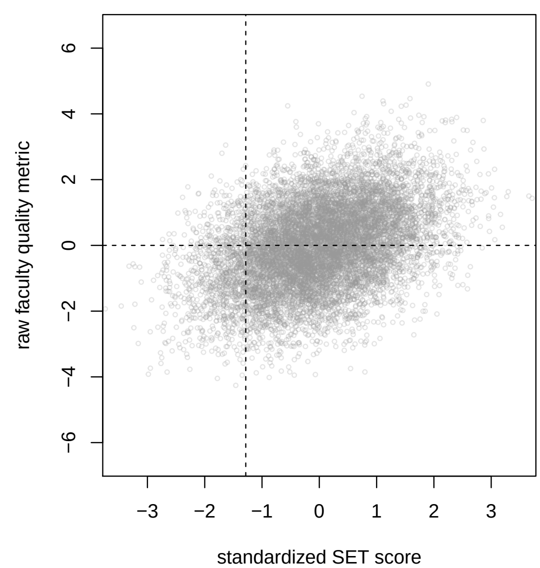
A simple example of this phenomenon is shown in Figure 6. Both panels of the figure depict a relationship between SET scores on the x-axis and faculty quality score on the y-axis. In both cases, SET scores are standardized with a mean of 0 and a standard deviation of 1; both panels also depict SET scores that are correlated with faculty quality at an identical $\rho \approx 0.4$. However, the left panel (Figure 6a) shows a bimodal distribution of SET scores and faculty quality: most faculty are distributed around a slightly above-average SET score and are widely dispersed in true quality, but a small number of faculty are reliably low quality instructors and also tend to score poorly on SETs. The right panel shows a bivariate normal (unimodal) distribution with the same variance-covariance matrix as the distribution in Figure 6a. The vertical dashed line shows the 10th percentile of SET scores for each case, while the horizontal dashed line shows the median faculty quality score.

In the bimodal distribution of Figure 6a, fewer than 1% of faculty below the 10th percentile of SET scores are better than the median instructor. Thus, using this percentile as a cutoff to identify poor teachers would be quite efficient and mostly fair in this scenario. However, in the unimodal distribution of Figure 6b, over 21% of faculty below the 10th percentile of SET scores are nevertheless above the median in terms of true quality. Thus, in this other case, using low SET scores to identify bad instructors would be unfair. Because we generally cannot directly observe the distribution of true faculty quality in any relevant population, a user of SET scores has no way to know which situation they are in. We think this demonstration underscores the need for caution, and reinforces our conclusion that SET scores are best used as an indicator of the need for a more thorough, costly, and accurate investigation of teaching performance.

Figure 6: **SET/Quality Relationship's Sensitivity to Distribution:** each figure shows a simulated relationship between 10,000 standardized SET scores (on the x-axis) and a true faculty quality metric (on the y-axis) in raw terms, without conversion to percentiles. In both panels, the correlation between SET scores and faculty quality is $\rho \approx 0.4$. The dashed vertical line represents the 10th percentile of SET score, while the dashed horizontal line represents the median (50th percentile) faculty quality score. The data are simulated using the `mvtnorm` library (Genz et al., 2018) in Microsoft R Open 3.5.3 (R Core Team, 2019).



(a) relationship between SET scores and faculty quality, bimodal distribution, $\rho \approx 0.4$



(b) relationship between SET scores and faculty quality, unimodal distribution, $\rho \approx 0.4$

References

Andersen, K. and E. D. Miller (1997). Gender and student evaluations of teaching. *PS: Political Science & Politics 30*(2), 216–219.

Benton, S. L., D. Duchon, and W. H. Pallett (2013). Validity of student self-reported ratings of learning. *Assessment & Evaluation in Higher Education 38*(4), 377–388.

Benton, S. L. and D. Li (2017). Idea student ratings of instruction and rsvp. Idea Paper No. 66. URL: `https://www.ideaedu.org/Research/IDEA-Papers-Series/Paper-Details?story=idea-student-ratings-of-instruction-and-rsvp` accessed 9/30/2019.

Benton, S. L. and S. Young (2018, June). Best practices in the evaluation of teaching. IDEA Paper No. 69. URL: `https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/IDEA_Paper_69.pdf` accessed 10/4/2019.

Boysen, G. A., T. J. Kelly, H. N. Raesly, and R. W. Casner (2014). The (mis) interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education 39*(6), 641–656.

Cascio, W. F. (1980). Responding to the demand for accountability: A critical analysis of three utility models. *Organizational Behavior and Human Performance 25*(1), 32–45.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research 51*(3), 281–309.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education 30*(6), 583–645.

Franco, A., N. Malhotra, and G. Simonovits (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science 345*(6203), 1502–1505.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2018). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-8. URL: `https://CRAN.R-project.org/package=mvtnorm`.

Hofert, M. (2018). *Elements of Copula Modeling with R*. Springer.

Hofert, M., I. Kojadinovic, M. Maechler, and J. Yan (2017). *copula: Multivariate Dependence with Copulas*. R package version 0.999-18, URL: `https://CRAN.R-project.org/package=copula`.

Kojadinovic, I. and J. Yan (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software 34*(9), 1–20. URL: `http://www.jstatsoft.org/v34/i09/`.

Langbein, L. I. (1994). The validity of student evaluations of teaching. *PS: Political Science & Politics 27*(3), 545–553.

Li, D., S. L. Benton, R. Brown, P. Sullivan, and K. R. Ryalls (2016). Analysis of idea student ratings of instruction system 2015 pilot data. *IDEA Technical Report No. 19*. URL: `https://www.ideaedu.org/Portals/0/Uploads/Documents/Technical-Reports/IDEA_Technical_Report_No_19.pdf` accessed 10/1/2019.

Marsh, H. W. (1982). Seeq: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology 52*(1), 77–95.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology 76*(5), 707–754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research 11*(3), 253–388.

Marsh, H. W. and L. A. Roche (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist 52*(11), 1187–1197.

Martin, L. L. (2016). Gender, teaching evaluations, and professional success in political science. *PS: Political Science & Politics 49*(2), 313–319.

Mengel, F., J. Sauermann, and U. Zölitz (2018). Gender bias in teaching evaluations. *Journal of the European Economic Association 17*(2), 535–566.

Mitchell, K. M. and J. Martin (2018). Gender bias in student evaluations. *PS: Political Science & Politics 51*(3), 648–652.

Nargundkar, S. and M. Shrikhande (2012). An empirical investigation of student evaluations of instruction–the relative importance of factors. *Decision Sciences Journal of Innovative Education 10*(1), 117–135.

Nargundkar, S. and M. Shrikhande (2014). Norming of student evaluations of instruction: Impact of noninstructional factors. *Decision Sciences Journal of Innovative Education 12*(1), 55–72.

Naylor, J. C. and L. C. Shine (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology 3*, 33–42.

Owen, D. and L. Li (1980). The use of cutting scores in selection procedures. *Journal of Educational Statistics 5*(2), 157–168.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. R version 3.5.3, URL: `https://www.R-project.org/` and `https://mran.microsoft.com/open`.

Schwarzer, G., J. R. Carpenter, and G. Rücker (2015). *Meta-Analysis with R*. UseR! Springer.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association 54*(285), 30–34.

Sterling, T. D., W. L. Rosenbaum, and J. J. Weinkam (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician 49*(1), 108–112.

Taylor, H. C. and J. T. Russell (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology 23*(5), 565–578.

Uttl, B., C. A. White, and D. W. Gonzalez (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation 54*, 22–42.

Online Appendix

Figure 7: **Distribution of True Instructor Quality among the Professors with Exceptionally High SET Scores:** the histogram depicts the distribution of a faculty member's true quality as an instructor (relative to the full population of all faculty members) if that faculty member has a SET overall instructor score above the 95th percentile. The bins are spaced such that the height of the bin represents the proportion of faculty members in the bin. Results are based from 1,000,000 draws from the normal copula with correlation between SET score and instructor quality fixed at $\rho = 0.4$. Over 18% of faculty members above the 95th percentile on SET scores are at or below the median on teacher quality.

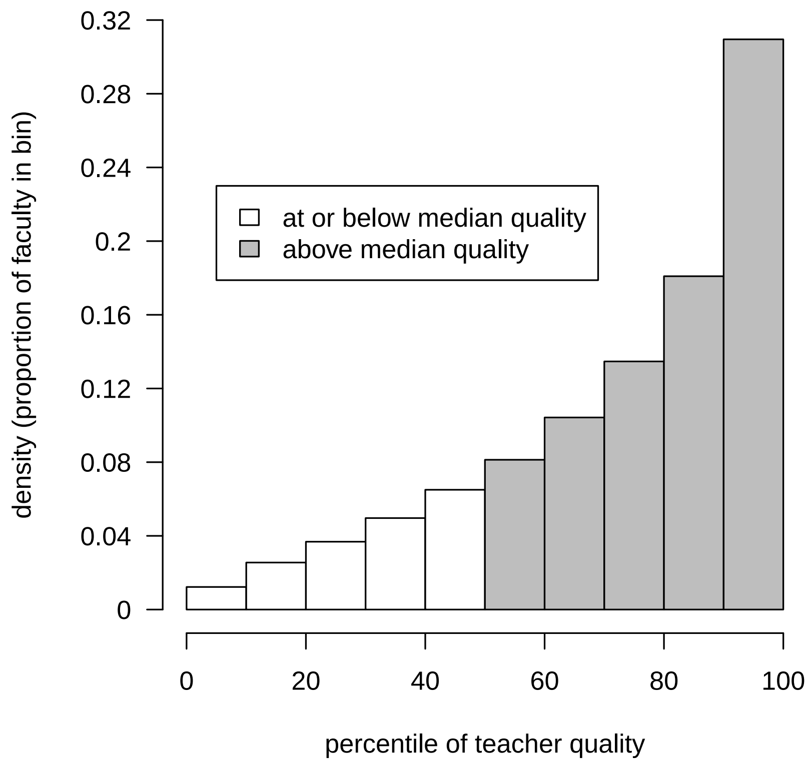


Figure 8: **Simulated SET Scores and Instructor Quality Levels, Normal Distribution:** the graphs shows the relationship between simulated class averaged SET scores (s_i, shown on the x-axis) and instructor quality ratings (q_i, shown on the y-axis) from a bivariate unit normal distribution with correlation $\rho \in \{0, 0.4, 0.9\}$; this repeats the analysis in Figure 1 with a bivariate normal distribution between SET scores and instructor quality. Both variables are modelled as standardized z-scores. Each point in the scatter plot (s_i, q_i) represents a simulated faculty member. Figure 1a shows an example with no correlation between SET score and instructor quality. Figure 1b shows an example with moderate correlation ($\rho = 0.4$) between SET score and instructor quality. Figure 1c shows an example with extremely high correlation ($\rho = 0.9$) between SET score and instructor quality. Each figure shows 3,000 simulated faculty members. The histograms on the top and right edges of each scatter plot show the normal distribution for SET scores (on the top edge) and instructor quality (on the right edge). The data are simulated using the `mvtnorm` library (Genz et al., 2018) in Microsoft R Open 3.5.3 (R Core Team, 2019).

(a) Correlation = 0 (b) Correlation = 0.4 (c) Correlation = 0.9

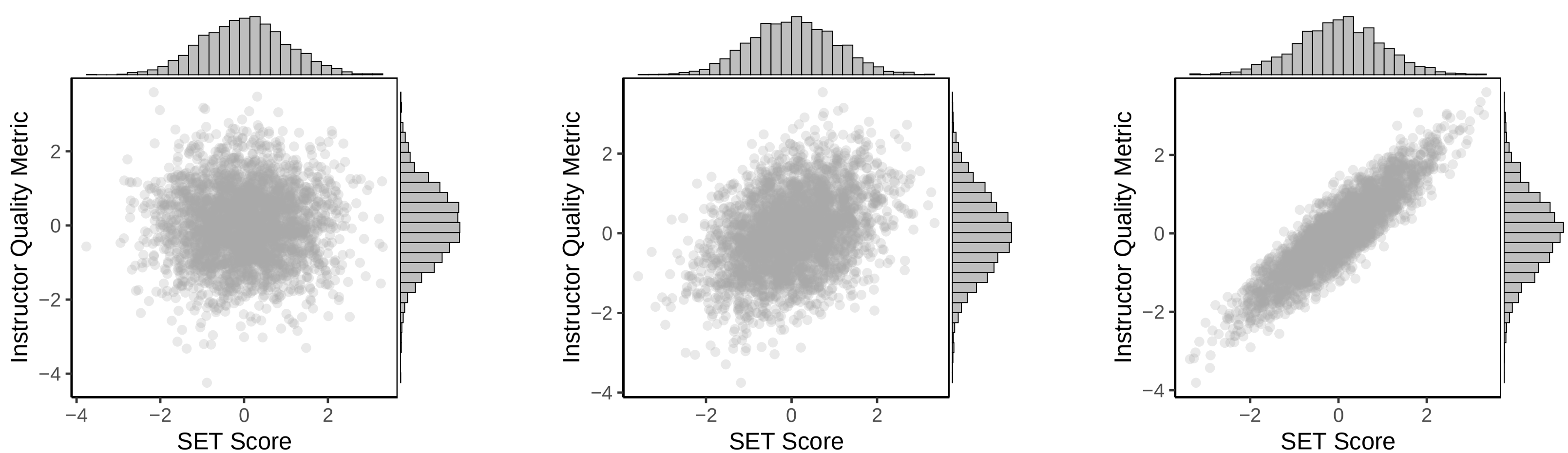


Figure 9: **Pairwise Comparison Error Rate, Normal Distribution:** the graph shows the ability of SET overall instructor scores to identify the better teacher in a pairwise comparison; this repeats the analysis in Figure 2 with a bivariate normal distribution between SET scores and instructor quality. The x-axis shows ρ, the correlation between SET overall instructor score s_i and actual instructor quality q_i as drawn from a bivariate normal distribution with mean 0 and standard deviation of 1. The y-axis indicates the estimated proportion of the time that the higher quality instructor also has the lower average evaluation (that is, $s_i - s_j$ and $q_i - q_j$ do not have the same sign). Each point in the plot represents an estimate from 500,000 simulated comparisons. Gray dashed lines indicate the simulated error rate when correlation is 0.2 and 0.9, as discussed in the text.

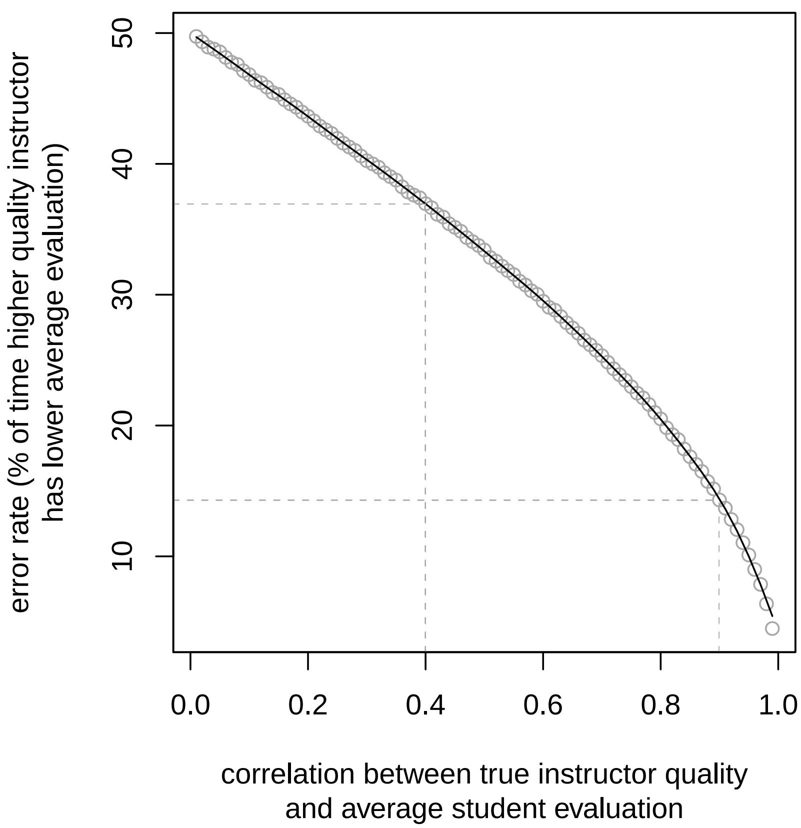


Figure 10: **Pairwise Comparison Error Rate with Minimum Gap, Normal Distribution:** the graph shows the ability of SET overall instructor scores to identify the better teacher in a pairwise comparison when scores must be at least a certain distance apart in order to conclude that the difference is meaningful; this repeats the analysis in Figure 3 with a bivariate normal distribution between SET scores and instructor quality. The x-axis shows the minimum distance between SET scores $g_{ij} = (s_i - s_j)$ that must exist before we decide that faculty member i is a better teacher than faculty member j. The y-axis indicates the estimated proportion of the time that the higher quality instructor also has a negative SET score gap larger than the minimum decision value. The correlation between SET overall instructor score and actual instructor quality ρ is fixed at 0.4. Each point in the plot represents an estimate from 500,000 simulated comparisons. Gray dashed lines indicate the approximate minimum difference in evaluation scores needed to achieve 10% and 5% error rates in decision-making.

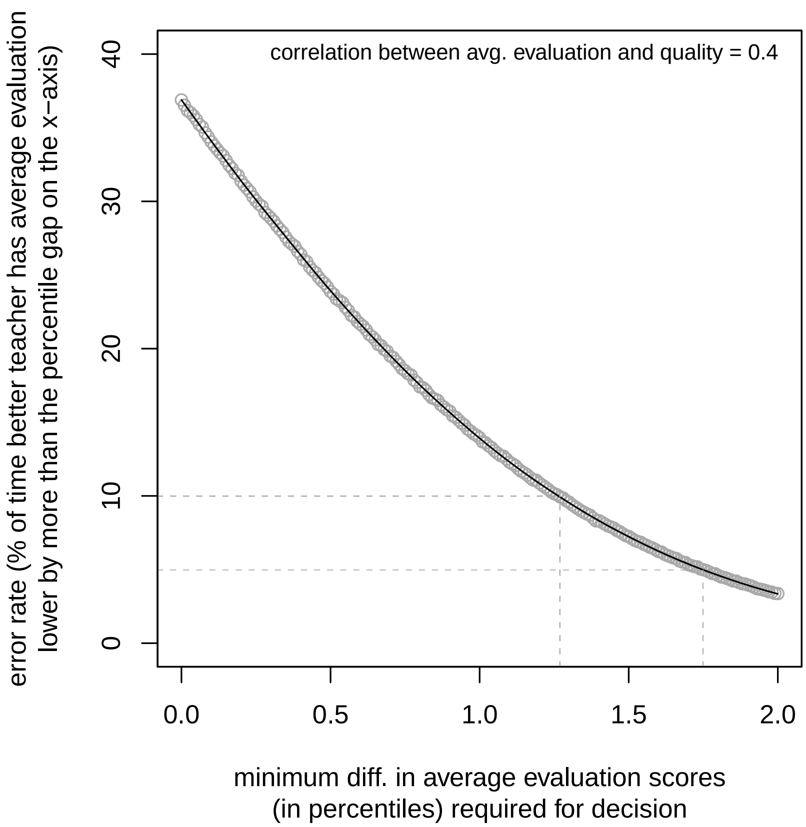


Figure 11: **Distribution of True Instructor Quality among the Professors with the Poorest SET Scores, Normal Distribution:** the histogram depicts the distribution of a faculty member's true quality as an instructor (relative to the full population of all faculty members) if that faculty member has a SET overall instructor score at or below the 20th percentile; this repeats the analysis of Figure 4 with a bivariate normal distribution between SET scores and instructor quality. Results are based from 1,000,000 draws from the bivariate normal density with mean 0, standard deviation of 1, and correlation between SET score and instructor quality fixed at $\rho = 0.4$. Just over 27% of faculty members below the 20th percentile of SET scores are above the median in teacher quality.

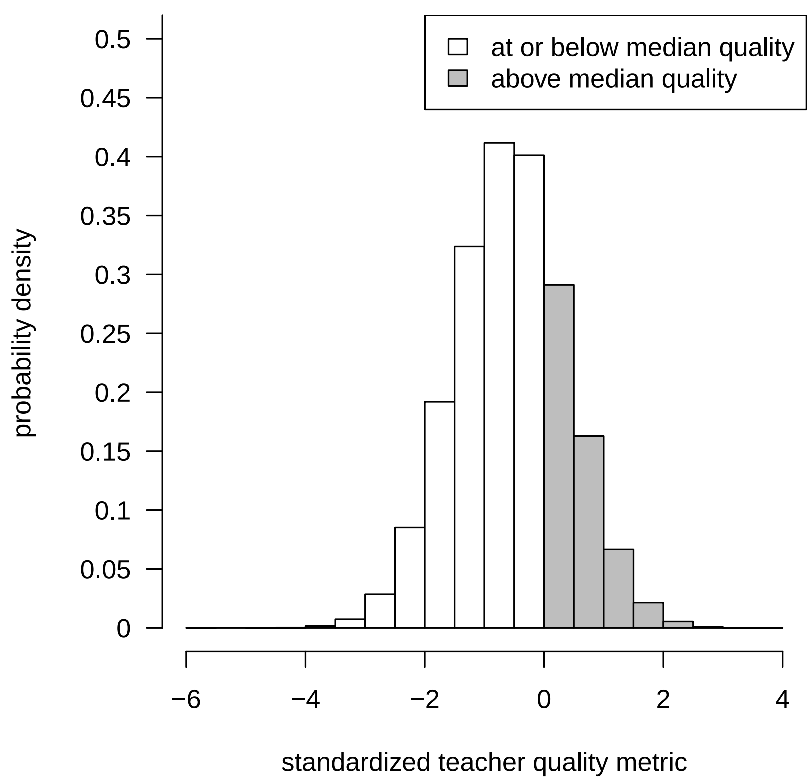


Figure 12: **Distribution of True Instructor Quality among the Professors with Exceptionally High SET Scores, Normal Distribution:** the histogram depicts the distribution of a faculty member's true quality as an instructor (relative to the full population of all faculty members) if that faculty member has a SET overall instructor score above the 95th percentile; this repeats the analysis of Figure 7 with a bivariate normal distribution between SET scores and instructor quality. Results are based from 1,000,000 draws from the bivariate normal density with mean 0, standard deviation of 1, and correlation between SET score and instructor quality fixed at $\rho = 0.4$. Over 18% of faculty members above the 95th percentile of SET scores are at or below the median in teacher quality.

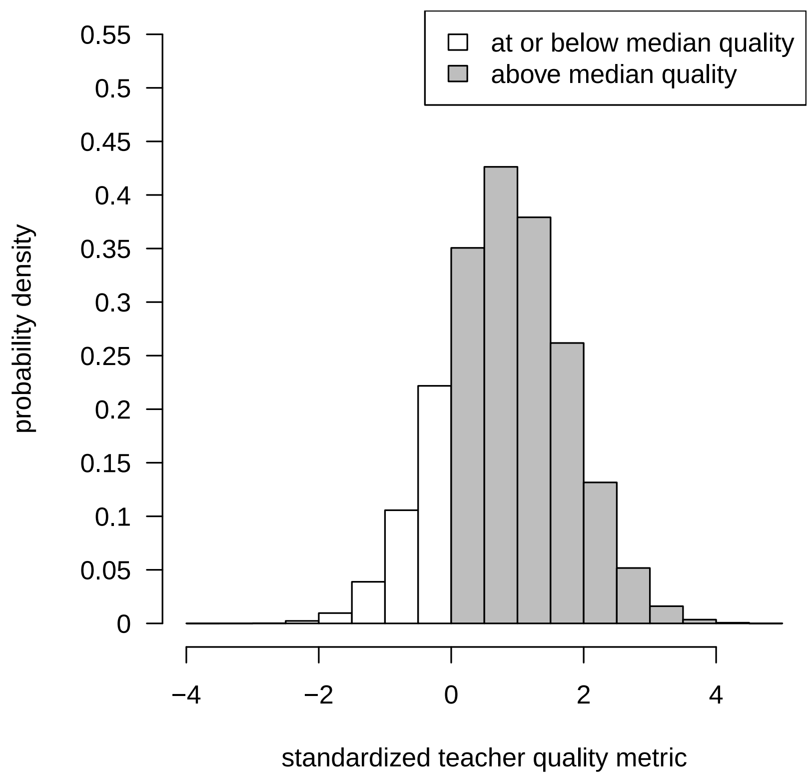


Figure 13: **Improvement in Measure Quality via Averaging, Normal Distribution:** the chart shows an estimate of the correlation between the average of multiple measures of instructor quality (on the y-axis) as a function of the degree to which the idiosyncratic, instructionally irrelevant influences on each measure are correlated with one another (on the x-axis); this repeats the analysis of Figure 5 with a multivariate normal distribution between instructor quality and four noisy measures. Correlation of the average measure is estimated using 10,000 simulations. In each simulation, a thousand observations of instructor quality and four imperfect measures of quality are drawn from a multivariate normal distribution with mean zero and standard deviation of one. Each measure is correlated with instructor quality at $\rho = 0.4$, and each measure is correlated with the other measures according to the collinearity (at equally spaced intervals between 0.1 and 0.9) indicated on the x-axis. The measures are averaged and the correlation of the average with true quality is calculated for all 10,000 simulations. The mean correlation across simulations is listed on the plot. The process is repeated using all four measures, only three measures, and only two measures; each is listed on the plot.