Essays

Analyzing Student Evaluations of Teaching: A Generic Prescription

Alexis Teagarden and Michael Carlozzi

Responding to calls for better use of student evaluations of teaching (SET) data, we report on a "generic" method of SET analysis. To test its efficacy, we generated score distributions from ten semesters of first-year writing course SET data in terms of unacceptable, adequate, and exceptional rankings by using three statistically orthodox approaches of categorizing scores and two versions of our generic method. We found that all methods yielded practically identical results. Our findings suggest WPAs have options for fair, transparent, and efficient use of SET data that do not require deep statistical expertise. More generally, we argue that if WPAs can promote responsible and sound methods of assessing SETs, they would not only improve the fairness of faculty evaluation processes but also help (re)establish themselves as critical voices in how such reviews should run. The complexity and copiousness of SET debates afford WPAs the opportunity to make such proposals since, we argue, SETs, like medicine or rhetoric, have value not in themselves but rather in their use.

Though scholarly debate swirls around student evaluations of teaching (SETs), several consensus points stand out. One is that the semesterly ratings of instructors play an increasingly prominent role in faculty evaluation worldwide (d'Apollonia & Abrami, 1997; Beran, Violato, & Kline, 2007; Linse, 2017; Wooten, Ray, & Babb, 2016). A second is that the typical SET form, with its mix of Likert-scale questions and open-ended comments, produces data that are difficult to analyze and interpret (Brockx, Van Roy, & Mortelmans, 2012; Darby, 2008; Dayton, 2015b; Gravestock & Gregor-Greenleaf, 2008; Harpe, 2015; Sullivan & Artino, 2013). Accordingly, a third consensus point arises: administrators and evaluators struggle to use SETs fairly in high-stakes decisions about faculty retention, promotion, and

tenure (Beran et al., 2007; Boysen, Kelly, Raesly, & Casner, 2013; Franklin & Theall, 1990; Linse, 2017; Thorne, 1980).

Many issues surrounding SETs can therefore be traced to not what they are but rather how they are used. For example, even SETs' strongest advocates argue these data should play a limited role in faculty evaluations, as they are "crude" measures (d'Apollonia & Abrami, 1997) and cover only one of teaching's many dimensions (Marsh, 2007). But reports show annual teaching reviews too often depend entirely on SET results (e.g., Franklin & Theall, 1990). Even within the field of Writing Studies, with its commitment to holistic, situated assessment, Moore (2015) claimed faculty evaluators "struggle themselves to match theory with practice when placed in a supervisory role" (p. 135). Wooten et al.'s 2016 field survey on SETs further demonstrated this problem, with WPAs reporting SET use to be prevalent but contested. The constant attention to SETs has not yielded widespread improvement in their use.

In considering how to improve faculty evaluation, Moore (2015) advocated practical responses to real constraints, including those of time and institutional demands. Dayton (2015b) proposed several best practices for handling these data, such as avoiding "norm referenced" evaluations, drawing on multiple forms of teaching evidence, and circulating a "written policy" regarding SET administration (pp. 41–42). Wooten et al. (2016) also proffered a set of guidelines for SET use, which emphasized consistency and transparency. Together these recommendations codify the principles that should guide SET use in faculty evaluation. But the articles do not go so far as to offer concrete measures for enacting these principles.

In response, our article offers a method for operationalizing the goal of consistent, fair, and transparent SET analysis as part of a wider faculty evaluation process. We do so by reporting on a "generic" SET review process that we developed for high-stakes, summative evaluation in a first-year writing program, using data that Wooten et al. (2016) found to be commonly made available to WPAs: means and standard deviations. Similarly, our generic method requires only a basic understanding of statistics and Microsoft Excel, and it aligns with the few points of consensus in SET literature, namely that SET data should be understood as permitting only broad evaluative characterizations such as "unacceptable, adequate, or exceptional" (d'Apollonia & Abrami, 1997, p. 1205).

While our method conforms to some SET research best practices, it deviates from traditionally prescribed methods of statistical analysis. So to test the efficacy of our generic method, we compared the results from two versions of our method to the results of three, statistically orthodox methods recommended in SET literature. For data we used ten semesters

of a first-year writing program's SETs. We found high to perfect agreement among all of the tested methods—in other words, the faculty scores almost always fell within the same evaluation category regardless of the method used. Since our generic method produced almost or entirely the same results as more resource-demanding ones, we argue that it should be regarded as a viable approach.

We conclude by stepping back from the method itself in order to discuss how SET analysis is inherently rhetorical work and how SETs are better understood as a rhetorical, rather than statistical, problem. In doing so, we discuss the pros and cons of the various methods reviewed in terms of a writing program's potential goals for faculty evaluation. For example, Wooten et al. raised questions about the role SETs can and should play in determining teaching excellence. For WPAs who shape SET review policies, we discuss ways of building analysis processes to either identify excellent instructors or to concentrate instead on delineating acceptable from concerning results.

For WPAs lacking such direct control over SET review, we argue our method could be used for internal assessment to help explain results to faculty and to direct coaching or mentoring discussions. Wooten et al. (2016) noted how WPAs are frequently assigned such roles and lack ways to make sense of SET scores for faculty (pp. 54–55); our generic method provides a simple and quick way to show faculty how to read scores according to SET scholarship's best practices, providing both transparency and the one-on-one consultation work often needed to make SET results useful to instructors (Boysen, 2016b; Neumann, 2000; Penny & Coe, 2004).

Maintaining consistency and transparency in high-stakes decision making is a constant good; it might also be a constant fight. But Wooten et al. (2016) also suggested that the contested role of SETs provides an opportunity for WPAs to establish authority and agency; we can see them as a site for what Adler-Kassner (2008) named "strategic action," or the harnessing of ideals and strategies. Adler-Kassner argued that WPA work has historically demonstrated the potential and significance of strategic action around two issues: assessment and labor. If we broaden assessment from its roots in student work to that of faculty, we can see how SETs create a space where assessment and labor issues meet (see also Dayton 2015a). Further, as Moore argued, given writing studies' historical engagement with assessment practices, WPAs "are in an ideal position to assist with campus-wide rethinking of faculty evaluation practices" (p. 147). Drawing on our disciplinary expertise and Adler-Kassner's strategic agency practices, WPAs can be important voices on improving the use of SETs and the overall evaluation of faculty.

Thus, in offering a method of efficient SET analysis, we also seek to intervene in larger issues regarding writing faculty and WPAs in particular. Overall, we join previous calls encouraging all WPAs to campaign for the ethical and effective use of SET data, to better support their program instructors, and to build their own ethos as experts in all aspects of faculty evaluation. Regardless of the exact form, we advocate a process that enacts fairness and transparency, aligns with local faculty evaluation priorities, and best allocates the often-scarce resources of time and skill—with Moore (2015), we argue those constraints are too significant to ignore. In this way, we argue WPA's subject matter expertise is necessary for understanding the contextual and disciplinary features of "unacceptable" SET results and for making decisions about concerning cases, and that SETs create a rhetorical problem, one not by solved statistical software but rather continually managed though situated knowledge and prudent judgment.

A Generic Prescription: Just What the Dean Ordered

Co-author Teagarden's experience with SETs aligned with many elements discussed in Wooten et al.'s (2016) survey; it differed in one key way. From her first year as WPA, Teagarden was granted a great deal of authority over SET evaluation, at least for the writing program's full-time and part-time instructors. In her first semester, she was tasked with independently analyzing SET scores for thirty-some faculty as part of the contractually mandated annual review, a now permanent responsibility.

As she began reviewing SET reports, Teagarden realized the project required not just analysis but also the creation of an entire evaluation procedure; her institution lacked formal guidance. What few protocols the author could find resembled what Wooten et al. (2016) described with rightful concern: instructor scores were compared to some mean, exemplified as "higher than average is excellent, within a couple decimal tenths is fine, [and] lower is concerning" (p. 58). Wooten et al. (2016) argued against such an acontextual emphasis on numerical results. In doing so, they echo a call long cried by the SET literature.

One of the few points of consensus within the highly-debated world of SET research is that administrators and evaluators struggle to make fair use of SET in high-stakes decisions about faculty retention, promotion, and tenure (Beran et al., 2007; Boysen et al., 2013; Franklin & Theall, 1990; Linse, 2017; Thorne, 1980). As early as 1980, Thorne was arguing that poor or absent methods for using SETs created serious issues: "we have rarely found explicit decision-making rules for the use of such data, so their potential administrative abuse has been omnipresent" (p. 214). While

Thorne ultimately reported on positive outcomes regarding SET use in faculty evaluation, research in the following decades grew grim. Aleamoni (1999) argued "the disadvantages of gathering student ratings primarily result from how they are misinterpreted and misused. The most common misuse is to report raw numerical results and written comments assuming that the user is qualified to interpret such results" (p. 160). More than fifteen years later, Boysen (2016a) discovered that faculty continued to misinterpret SET data, irrespective of their statistical training. Reading the last forty years of research on the use of SETs leads one to conclude these data are used everywhere and everywhere used badly.

Even those who champion SET usage have registered alarm at how SETs are interpreted. Gravestock and Gregor-Greenleaf's (2008) comprehensive review of the literature calls attention to "a significant absence of policies regarding, or information available to instructors and administrators providing guidance about the interpretation of course evaluation results [. . .] most institutional policies and information address only the process of conducting evaluations and disseminating the results" (p. 18). Decision makers have also been found to employ flawed metrics for evaluating SETs (Boysen et al., 2013) and to base important decisions on questionable comparisons (Franklin & Theall, 1990). While one might have expected the rising use of SET to bring better methods, the opposite appears true. As the significance of SET data expanded, concerns about their misuse also proliferated (Dewar, 2011; Linse, 2017; Palmer, 2012).

Poor interpretative practices deserve scrutiny, but they also warrant sympathy. For multiple reasons, SET data resist simple analysis. We summarize four key reasons below, since, understanding these reasons can help WPAs build better review processes and promote them to cross-disciplinary audiences.

First, most SET data come from Likert scales (Gravestock & Gregor-Greenleaf, 2008). Statisticians continue to debate how to best analyze Likert data (Harpe, 2015; Sullivan & Artino, 2013). Meanwhile, research on how to analyze open-ended SET comments is nascent, further complicating evaluation methods and prompting more debate than policy (Brockx et al., 2012), though Wooten et al. (2016) suggested that when WPAs have access to comments, they feel well-prepared to handle such data.

As SET data are generated from Likert scales, it is unsurprising that they are non-normally distributed (Darby, 2008). Thus, many scholars warn that neither parametric tests (such as *t* tests) nor popular descriptive statistics (mean and standard deviation) are appropriate comparative measures (McCullough & Radson, 2011; Mitry & Smith, 2014). Non-normally distributed data generally require specific methods of analysis that may be

unfamiliar or unknown to evaluators. These alternative analyses furthermore require SET scores to be reported in specific ways or that evaluators are capable of transforming data to the needed form.

Likert data are also ordinal, further muddying interpretation. There is not usually a normed or objective standard for choosing one score over another. In the case of SETs, what one student may rate as "strongly disagree" another may rate only as "disagree" or even "neutral." This has led some authors to claim that means and standard deviations do not comprise a "valid metric" and thus use of them "should cease" (McCullough & Radson, 2011, p. 189). Such arguments emphasize how difficult it is to imagine a meaningful average between, for example, strongly disagree and strongly agree.

Finally, a significant challenge to using SETs in summative decisions comes from generating acceptable and unacceptable rankings. Any evaluation must include the possibility of finding a faculty member's scores below expectation. But at what point should an instructor's scores be judged acceptable or unacceptable? The answer will always be non-statistical and thus open to the charge of arbitrariness.

So what was an allowable way to analyze SETs, Teagarden wondered? Scholars did offer multiple solutions to the problem of evaluating SET data. Three common themes emerged:

1. Require statistical training for any faculty evaluator engaged in SET review (Emery, Kramer, & Tian, 2003).

2. Use appropriate methods for analyzing non-normally distributed data when comparing faculty, such as interquartile range (IQR) and median, interpolated median, or proportions (McCullough & Radson, 2011; Mitry & Smith, 2014).

3. Use null hypothesis significance testing, such as *t* tests, when comparing faculty to determine if SET scores significantly differ, given that such tests tend to remain effective even when data violate normality assumptions (Boysen, 2015).

As Teagarden reviewed potential analysis protocols, the published recommendations began to resemble the marketing of name-brand commercial drugs. Each new version promised an innovative solution, a novel delivery system, a more personalized approach. They offered much, and they cost more. Proposed solutions invariably required advanced statistical knowledge, specialized software, extensive data preparation, weeks of one-on-one discussions, or all of the above.

For example, mandating only trained experts to interpret SET would likely require dramatic shifts in institutional staffing; moreover, the efficacy of this solution is questionable, as some research suggests expertise cannot guarantee proper analysis (Boysen, 2016a). The second and third solutions entail significant time and labor costs. Calculating appropriate methods of comparisons, such as an IQR, require that SET data be reported in specific ways and that evaluators have the expertise to manage these comparisons and the hours needed to calculate them. Such conditions strike us as unlikely for many departments and institutions, not just writing programs.

Teagarden's institutional context did not support any—let alone all—of these costs.

But even in departments or programs backed with statistical expertise, time constraints and labor distribution raise barriers (Moore, 2015). Faculty evaluation often occurs at the end of the academic year, with little time between the distribution of SETs and the review process. If (or, more likely, when) departments and programs cannot find an accommodating expert, how are they to follow the institutional mandates for SET use and the literature's guidelines for interpreting them?

Perhaps institutions with deep pockets can absorb such a bill without disruption. For Teagarden's program, as for many we suspected, a new approach was necessary. And if published solutions were the branded drugs of SET analysis, we thought perhaps there were generic options available. Generic drugs cost much less while offering the same active pharmaceutical ingredient. A generic SET process, by analogy, would cost less in time and labor while offering the same results in faculty evaluation. Thus, we developed an SET analysis method that worked with the typical skillsets of WPAs and within their typical constraints. We aimed to create a method that would produce fair, transparent, and efficient SET data analysis so that WPAs could spend more time on cases the process flagged as unusual or concerning—cases that called for subject matter and local program expertise.

We based our method—cheaper in resource demands, easier to implement, and just as good in results—on consensus points of SET scholarship rather than those of statistical analysis. That is, we

- compared SET scores only among instructors within a "similar teaching context" (Marsh, 2007; Neumann, 2000);
- calculated a faculty member's aggregate mean within the similar teaching context (Boysen, 2015; Harrison, Douglas, & Burdsal, 2004);

- evaluated scores following the standard three-tier category system: "unacceptable, adequate, or exceptional" (d'Apollonia & Abrami, 1997, p. 1205).

Then we tested to see if our generic version would offer the same results as more traditional prescriptions.

Research Method and Data

In seeking out a generic method of SET analysis, we compared results from various interpretative methods. We applied all methods to SET data from a public, American northeastern, doctoral research university's first-year writing program: the sequence of fall English 101 courses and spring English 102 ones. We thus followed the best practice of comparing SET scores only among instructors within a "similar teaching context," which here meant a required, introductory-level writing course capped at twenty-five students (Neumann, 2000). Within each semester, instructors were compared only to those teaching the same course numbers. English 101 and English 102 could not reasonably be compared within the same semester because they differed in not only pedagogy but also in student population, class sizes, and the offers of additional tutoring support.

We requested the first-year writing program's past ten semesters of anonymized SET data from the institutional research department. We received 6,075 completed SETs for 247 total sections/instructors. SET questions were 1–5 scale Likert-type items responding to a statement in terms of agreement such as "The instructor was prepared for class." The scale was symmetrical, with 1 equaling "strongly disagree," 2 equaling "disagree," 3 equaling "no strong opinion," etc. Students responded to fifteen distinct statements about the course, rating concepts such as the instructor's availability, overall effectiveness, and preparedness.

Institutional research provided data in the form sent to faculty and evaluators: aggregated score tables with a section's calculated *N*, mean, median, and SD along with those for the department's overall *N*, mean, median, and SD at that course level, i.e. all 100-level courses. To conduct the literature's recommended analyses, the data required significant transformation. We were given almost 7,000 rows of data that reported only aggregate counts for each individual instructor; for example, we might learn that 13 students in one class responded 5 (strongly agree) on one specific question. Since we needed to rank data, Carlozzi coded a Python program to convert these aggregate counts into individual student responses and then exported them to a spreadsheet for further analysis. We then followed recommended prac-

tice by combining the instructor's semester of classes to create one aggregate mean (Boysen, 2015; Harrison et al., 2004).

For every semester of received data, each instructor was scored according to all of the aforementioned evaluative systems, receiving a label of "unacceptable, adequate, or exceptional," following the standard three-tier category system (d'Apollonia & Abrami 1997, p. 1205). Comparisons between methods were first made to determine agreement when rating instructors as adequate or unacceptable. Comparisons were made again to determine agreement on rating instructors as adequate or exceptional. Ratings were inclusive; instructors rated exceptional also received ratings of adequate.

When comparing instructors, we investigated five distinct analytical methods, three from the published literature to represent resource-demanding "name-brand" methods and two versions of our "generic" one.

Traditional SET analysis models:

1. IQR and median (IQR/Median). Student scores were ranked and then an acceptable range was generated based on the course's interquartile range. For every semester, the first quartile was 4 and the other quartiles were 5. Instructors whose median scores fell below 4 were rated unacceptable.

2. IQR and interpolated median (IQR/IM). As above, the first and second quartiles were 4 and 5. However, because this statistic interpolates scores, it afforded more opportunity to identify unacceptable results.

3. *T* test. Because of unequal variances and sample sizes between groups, Welch's *t* tests were run. Instructors who differed significantly from course means were rated unacceptable or exceptional, depending on the direction, and according to a significance level of 0.05. *T* tests were used because they are recommended in the literature; however, it should be noted that multiple comparisons increase the familywise error rate.

Our Generic SET analysis models:

1. M+One. Generic version A, with mean and one standard deviation. Instructors whose means fell below one SD of the course mean were deemed unacceptable; one SD above were deemed exceptional.

2. M+Half. Generic version B, with mean and one half of a standard deviation. As above, except that one half of a SD was used rather than one.

As part of these comparisons, we considered each method as an independent "rater" and calculated inter-rater reliability through Cohen's kappa coefficient. Although somewhat controversial over its calculation for random effect (Guggenmoos-Holzmann, 1993), kappa attempts to produce agreement coefficients by estimating those agreements between raters that may have occurred "by chance."

Low kappa values in spite of high agreement stem from the "kappa paradox" whereby kappa values are lower in datasets with a high prevalence index (Feinstein & Cicchetti, 1990). When this occurs, kappa assumes that a tremendous number of cases will agree by chance. Our data had exceptionally high prevalence indices; in almost all of our comparisons, kappa assumed that the methods should agree by chance over 90% of the time. That is because unacceptable scores were rare; most instructor scores fell into the adequate category. Therefore, a prevalence-adjusted bias-adjusted kappa (PABAK) statistic is also presented whereby the expected agreement by chance is held constant. Readers may interpret kappa and PABAK values as they see fit, but the literature suggests general—though arbitrary—guidelines: < 0.21 = slight agreement; 0.21–0.40 = fair agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = substantial agreement; 0.81–0.99 = almost perfect agreement.

Proportions (McCullough & Radson, 2011) were not included because they required too arbitrary a judgment for discernment. Median and interpolated median worked alongside interquartile range, for example. Proportions, on the other hand, required an arbitrary decision on what constituted an acceptable cut-off point, one we could not confidently make.

Results: What Comparisons of SET Analysis Methods Reveal

In Tables 1–3, Percentage Adequate Instructors is the total percentage of instructors that the statistic rated as adequate or exceptional. Agreement is the percentage of time when a mean differentiation method agreed with another method when rating instructors as at least adequate. Percentage Exceptional Instructors is the percentage when that method—not compared to any other—rated instructors as exceptional. Kappa and PABAK columns report their respective values.

First, we compared the generic methods to IQR/Median. IQR/Median and M+One were functionally identical, agreeing in all instances. They were, however, unable to discriminate among instructors, rating 99% of teachers as adequate. M+Half discriminated better in determining unacceptable instructors (92% adequate).

Table 1

Mean differentiation compared to IQR/Median

Method	Agreement With IQR/ Median	% Adequate Instructors	Kappa	PABAK	% Exceptional Instructors
IQR/Median	-	99%	-	-	0%
Mean + 1 SD	100%	99%	1	1	0%
Mean + Half SD	93%	92%	0.09	0.86	1.6%

Course *N* = 247; SET *N* = 6,075.

We then compared mean differentiation to IQR/IM. As expected, IQR/IM was more discriminating than IQR/Median, rating 93% of instructors as acceptable. Agreement for M+Half was higher than with IQR/Median because interpolation could identify low performing instructors (table 2).

Table 2

Mean differentiation compared to IQR/IM

Method	Agreement With IQR/IM	% Adequate Instructors	Kappa	PABAK	% Exceptional Instructors
Interpolated Median	-	93%	-	-	0%
Mean + 1 SD	94%	99%	0.12	0.89	0%
Mean + Half SD	97%	92%	0.82	0.95	1.6%

Course *N* = 247; SET *N* = 6,075.

We then compared mean differentiation to *t* tests. The *t* test had the greatest discernment, scoring the most instructors as unacceptable and exceptional. It agreed overwhelmingly with mean differentiation (table 3).

Table 3.

Mean differentiation compared to Welch's *t* test

Method	Agreement with Welch's *t* test	% Adequate Instructors	Kappa	PABAK	% Exceptional Instructors
Welch's *t* test	-	89%	-	-	20%
Mean + 1 SD	89%	99%	0.06	0.79	0%
Mean + Half SD	95%	92%	0.73	0.91	1.6%

Course $N = 247$; SET $N = 6{,}075$.

Discussion: Aligning Methods with Summative Evaluation Purposes

Our results suggest that nonstandard comparative measures may effectively analyze SET data. IQR/Median, one recommended methodology for analyzing non-normally distributed data, performed identically to M+One, agreeing 100% of the time.

But this perfect agreement lacks utility for summative evaluations, as neither method could differentiate among instructors; only one instructor out of 247 was rated as unacceptable. This conclusion might satisfy some evaluators, as it suggests that all SET scores meet expectations, but this is not a conclusion we are willing to draw. To illustrate our concern, consider the case of Instructor P and Instructor R, scores taken from the same semester. In the below charts, Instructor P and Instructor R received median scores of 4 and were rated adequate under IQR/Median and M+One. Instructor P's results look respectable, scoring only 11% unacceptable responses.

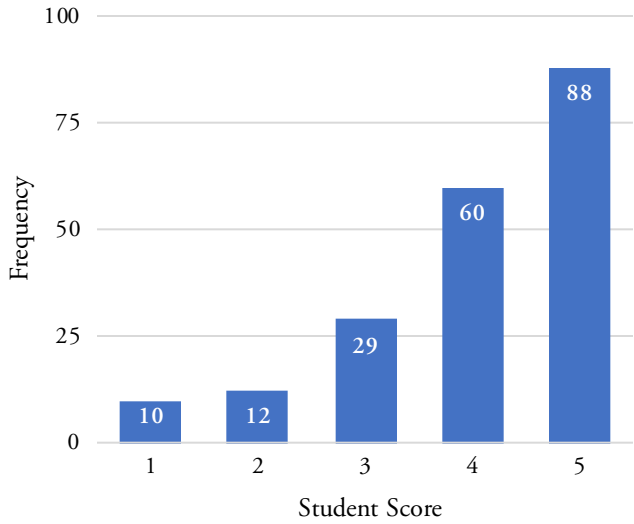

Figure 1. Instructor P's chart, fall 2015. Mean = 4.0; SD = 1.1; Median = 4.

Instructor R's results, however, are more problematic, with 28% unacceptable scores . This instructor was rated as unacceptable in M+Half, IQR/IM, and Welch's t test. We find this a more intuitive conclusion; an instructor receiving about 29% of responses as 1s and 2s should not be equated with one with only 11% of such responses.

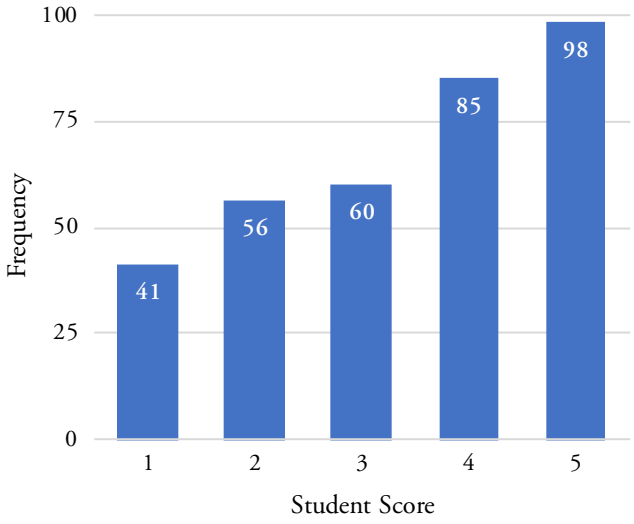

Figure 2. Instructor R's chart, fall 2015. Mean = 3.4; SD = 1.3; Median = 4.

This lack of differentiation suggests that those universities that have chosen interpolated median have done so for good reason. IQR/IM rated 7% of instructors as unacceptable, similar to results from other methods. But generating this result demands considerably more work and time, without any obvious advantage, over mean differentiation. Indeed, almost 98% of the time IQR/IM agreed with M+Half. In practical terms, they were essentially identical with respective kappa and PABAK values of .82 and .95. In other words, M+Half approximates the results of IQR/IM without the latter's logistical headaches. The data similarly show strong agreement among mean differentiation and the *t* test.

But efficient and statistically aligned analysis answers only some of the issues involved with SETs. Wooten et al. (2016) raised other discussion points to consider. They claimed "Focusing on SET averages alone is difficult to justify" (p. 58) and supported this by noting concern with the influence external factors have on SETs, such as "age, gender, level of course, and/or if course is required or elective" (p. 59). The best practice of comparing faculty's SET results among "similar teaching contexts" eliminates issues of course level and status; we believe every WPA should advocate for this approach and model it within their own reviews. Indeed, writing programs' hallmark of many-sectioned courses offers an ideal starting place for such practices.

The SET literature is more divided on how much age, gender, and race biases affect SET scores. Gravestock and Gregor-Greenleaf (2008) concluded: "In general, no variables have been found to have a substantial effect (e.g. something that would alter the ratings beyond the second decimal place) on ratings, except for expected grades" (p. 39). How to understand the "expected grade" variable remains contested. The field has developed several competing interpretations, with the "grade leniency hypothesis" and "validity hypothesis" predominating. While review of this debate is beyond the scope of our article, interested readers can consult Dayton (2015b) for a summary and Brockx, Spooren, and Mortelmans (2011) for a comprehensive treatment.

As the field of SET research is vast, articles can be found that support any number of positions, especially on the question of gender bias. Spooren, Brockx, and Mortelmans's (2013) review pointed out two articles showing female faculty received statistically significantly higher SET scores than male faculty and one showing the reverse. However, some of the most rigorous studies concluded that if gender bias exists, its effects are small enough to be eliminated by well-established analysis processes (Li & Benton, 2017). An appropriately wide "adequate" range could therefore wash out differences due to bias. Our generic model offers a further affordance. If a WPA

determines a particular external factor to severely bias results, then that factor can be considered part of the "similar teaching context." If gender bias is the concern, for instance, then gender could be a required element of the context, with faculty members only compared to instructors of the same gender. This is easier said than done, of course; the larger point is our generic method provides several ways to account for possible biasing factors.

More importantly, our argument for efficient data review aims to make time for the WPA to consider borderline cases or check for concerning patterns. Wooten et al. (2016) argued "Several reasons may explain an instructor's high or low numerical scores, and it is incumbent on WPAs to discover those reasons rather than risk false assumptions about someone's effectiveness based on numbers alone" (p. 59). This position matches that held by SET advocates: SETs should never operate as the only form of evidence for teaching review (Marsh, 2007).

Since SET data are likely a part of all faculty evaluation, careful review is necessary; as Moore (2015) argued, it is also a time-intensive task. One goal of our generic method is thus to allow a WPA to quickly distinguish between the majority of "adequate" cases and the few outliers, precisely so a WPA can conduct a deeper review and better determine the reason for unusual scores. We also hope the expedited process affords the WPA time to analyze overall results and identify concerning patterns by, for example, checking for systematically lower scores within a specific category like gender, race, rank, or age.

Wooten et al. (2016) also noted a concern about SETs' role in determining teaching excellence. For example, they argued "WPAs may want to openly question why [SETs] would be used to sanction some instructors and not used to commend others" (p. 61), but here we must disagree. Following d'Apollonia & Abrami (1997), we see SET scores as "crude" measures, unable to discern fine detail. We argue SETs should therefore have a limited role to play in determining teaching excellence. A pattern of consistently exceptional scores would point to a faculty member's ability to connect with and support students. While a writing program might deem that a necessary feature of teaching excellence, it cannot be sufficient. Rather, just as we argue "unacceptable" SET results call for further review to identify issues, "exceptional" scores call for consideration about what is working so well and why. It is in that discussion that multi-faceted evidence for teaching excellence can emerge. Thus, we argue for minimizing the role of SETs serve in either "sanctioning" or "commending" faculty. However, we also note that our generic method can be tailored to suit a program's goals. A WPA seeking ways of identifying outliers at both ends can narrow the range; one seeking to minimize exceptions can widen it.

Finally, Wooten et al. (2016) found a majority of survey respondents have mentoring roles attached to SET scores. We see our generic method as a means of navigating scores for faculty, even if the process cannot be used in formal assessment. Research has shown faculty struggle to make sense of SET scores, and active, engaging consultations are the best way to ensure that SET feedback improves teaching practice (Boysen, 2016b; Penny & Coe, 2004). Our generic method can help faculty put their results into a specific perspective—that of comparison across similar teaching contexts and within general categories. Teagarden has used this method with her campus's writing program faculty and finds it demystifies SET scores, to the relief of many and the disappointment of some. Approaching SET data this way shows how scores often mean less than they initially appear. Using our method for coaching can therefore afford opportunities to calm fears, but we caution WPAs that it can also puncture self-images, when faculty come to see that scoring above a mean does not automatically translate to an "exceptional" score. Avoiding "unacceptable" scores reassures many, but being called "adequate" can upset others. Thus, with the generic method, as with any other, discussion and contextualization are necessary to help faculty understand what terms mean and how to interpret data (Neumann, 2000; Penny & Coe, 2004).

Limitations

This study was limited by sample; we looked at one program in one university. And because we worked with anonymized data, we were unable to examine instructors longitudinally. Our program averaged around 30 instructors per semester; it did not have 247 distinct instructors over ten semesters. We are therefore unsure how these methods compare to each other if applied to dramatically larger sample sizes. Further research could examine how robust and congruent IQR and SD are around other kinds of data, in different institutions, and with programs other than first-year writing.

Our study also analyzed only one form of SET data, a composite mean score for all questions. SET scholarship remains divided on the best kind of data to generate and use. Each department will need to consider which data to analyze, be it a single "overall" question (e.g., "how effective was this instructor's teaching?"), a weighted formula of multiple questions, or, as our approach here, an instructor's semesterly composite mean. Additionally, as we drew on data from first-year writing courses, we followed the best practice of comparing faculty teaching within a "similar teaching context." Departments and institutions evaluating more heterogeneous teach-

ing contexts would need to perform additional work establishing reasonable comparators.

And, taking a wider view, we acknowledge the limitations inherent in SET data. Fair and effective use of SET data might be a necessary part of faculty evaluation, but it alone is not a sufficient representation of teaching's complex art. A larger challenge may be raised that SET data deserve no standing in faculty review processes, as they support problematic aspects of the university, such as neoliberal market rhetoric in general and, more particularly, contingent labor practices or student-as-consumer frames (Crowley, 1998; Schweitzer, 2009). We acknowledge this perspective but choose to advocate for strategic action over outright renunciation. Simply put, we believe students have important, if limited, insights into teaching and that instructor efficacy merits attention. There are better and worse ways of conducting SET analysis; we argue WPAs should take positions on how to best use SETs rather than reject them outright.

Finally, we reiterate that this analysis is strictly comparative. That is, it aims to identify agreement among analytical methods recommended in the literature to our generic method. The validity and appropriateness of the primary methods remain separate concerns. Some evaluators, for example, may protest about using *t* tests on skewed data. We stress that these issues are separate from the current analysis, and we guide interested readers to the robust literature on SET data (Marsh, 2007; Spooren et al., 2013).

Concluding Remarks on Statistical vs. Rhetorical Problems

In developing and testing a "generic" method of SET analysis, Teagarden drew, in part, from her upbringing. She is the daughter of a pharmacist; medical metaphors come naturally. But the metaphor of drugs also emphasizes the rhetorical nature of SETs and their use. To read SET articles for any time is to be reminded of Gorgias' comparison of speeches and drugs, where some "cause pain, some pleasure, some fear; some instil courage" (p. 287). SETs elicit the same range of responses, and the divergence can often be traced back to how fairly and transparently these data are treated.

For as with rhetoric, the analogy to drugs reminds us that SETs' value is not inherent but rather emerges from their use. Almost all SET advocates argue they provide only a rough measure of a single teaching facet. We agree and argue this simple sorting is an important first step—not a final one. We also argue that impossible-to-implement recommendations serve no one. Moore (2015) enumerates the many hurdles departments face when trying to perform multi-faceted evaluation of faculty. Tight deadlines cannot justify unfair assessment, of course, but SET analysis must work

within the likely constraints administrators and evaluators face. Time limits are a real factor; limited statistical expertise is another. These are particularly likely to affect evaluators' ability to use IQR, IM, and significance testing; the same is doubly true for more complicated measures like the "distributions of responses" (p. 61) referenced by Wooten et al. (2016). For instance, obtaining and analyzing raw student responses proved challenging for us, even though we explicitly requested the data in a specific format and worked outside of customary evaluation timelines. To render data useable for our analysis also required considerable labor as well as computer programming skills. What hindered our work could completely stop other evaluators working with fewer resources or different training.

Such issues cannot be underestimated. As a case study, Samuels (2018) recounts how his institution's team was unable to alter the role of SETs in non-tenure-track (NTT) faculty review, in part because "university administration told us that it would be too costly and time-consuming to develop a different model of performance evaluation" (p. A23). While we do not entirely share all of Samuels's views towards SETs, we do agree that SET use should conform to ethical principles, and that if such processes are to win institutional approval, they must work within local constraints. As Samuels's case illustrates, evaluation systems perceived as requiring too many resources can be rejected out of hand. But since our generic approach showed minimal variation in results from those generated by the more resource-demanding methods, we argue the advanced recommendations in the literature are unnecessary. Our mean differentiation approach—easily generated and distributed—achieves the same goal.

To conclude, for our specific dataset, we found that mean differentiation (mean + a chosen SD range) provided an efficient way to compare faculty scores, given one works within the research consensus on similar teaching contexts and differentiation (e.g., exceptional, adequate, and unacceptable). Agreement was strong among all methods tested. We suggest WPAs seeking ways to best use SET in summative evaluations, and with similar SET score distributions as ours, adopt one of proposed mean differentiation methods for summative or formative purposes, as their local context allows. Evaluators can be confident that our generic method, while perhaps not in harmony with orthodoxy, has been supported by empirical data.

While the data analyses work out similarly, we also note such results call for careful interpretation. Recognizing this, we also argue evaluators use these findings as "alerts" rather than immediately act on them. One benefit to efficiently generated analysis is that it allows evaluators to quickly sort the standout cases from the unexceptional ones. Having completed a general delineation, evaluators can devote more attention to borderline

cases, determining which cases require further information before rendering a judgement.

As such our method illustrates how SETs must be understood as rhetorical, not merely statistical, problems. To treat them as objective data is to misread what numerical ratings report, to mistake teaching for a single-faceted activity to, and to miss entirely the real work involved with evaluating faculty. Dayton (2015b), however, framed this persistent problem as a rhetorical opportunity, arguing that writing programs can address these issues by making SET analysis more transparent to all stakeholders, including faculty members, students, and "the larger constituencies who are nudging us in this direction" (p. 42). In other words, if how an institution treats SETs mirrors the way it treats WPAs and writing faculty in general, then creating fair and transparent approaches to these data can help further situate WPAs and writing instructors as authoritative experts who responsibly engage with institutional questions of education quality.

Acknowledgments

We would like to thank our peer reviewers for their constructive feedback and J. Russell Teagarden for his inspired metaphors.

References

Adams Wooten, Courtney, Ray, Brian, & Babb, Jacob. (2016). WPAs reading SETs: Toward an ethical and effective use of teaching evaluations. *WPA: Writing Program Administration, 40*(1), 50–66. Retrieved from http://www.wpacouncil.org/wpa40n1

Adler-Kassner, Linda. (2008). *The activist WPA: Changing stories about writing and writers*. Logan, UT: Utah State University Press.

Aleamoni, Lawrence M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153–166. doi:10.1023/A:1008168421283

Beran, Tanya, Violato, Claudio, & Kline, Don. (2007). What's the 'use' of student ratings of instruction for administrators? One university's experience. *The Canadian Journal of Higher Education*, *37*(1), 27–43. Retrieved from https://eric.ed.gov/?id=EJ771048

Boysen, Guy A. (2015). Preventing the overinterpretation of small mean differences in student evaluations of teaching: An evaluation of warning effectiveness. *Scholarship of Teaching and Learning in Psychology, 1*(4), 269–282. doi:10.1037/stl0000042

Boysen, Guy A. (2016a). Statistical knowledge and the over-interpretation of student evaluations of teaching. *Assessment & Evaluation in Higher Education, 41*, 1–8. doi:10.1080/02602938.2016.1227958

Boysen, Guy A. (2016b). Using student evaluations to improve teaching: Evidence-based recommendations. *Scholarship of Teaching and Learning in Psychology, 2*(4), 273–284. doi:10.1037/stl0000069

Boysen, Guy A., Kelly, Timothy J., Raesly, Holly N., & Casner, Robert W. (2013). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment and Evaluation in Higher Education, 39*(6), 641–656. doi:10.1080/02602938.2013.860950

Brockx, Bert, Spooren, Pieter, & Mortelmans, Dimitri. (2011). Taking the grading leniency story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. *Educational Assessment, Evaluation and Accountability 23,* 289–306. doi:10.1007/s11092-011-9126-2

Brockx, Bert, Van Roy, Karlijn, & Mortelmans, Dimitri. (2012). The student as a commentator: students' comments in student evaluations of teaching. *Procedia Social and Behavioral Sciences*, *69*, 1122–1133. doi:10.1016/j.sbspro.2012.12.042

Crowley, Sharon. (1998). The politics of composition. *Composition in the University: Historical and Polemical Essays*. Pittsburgh, PA: University of Pittsburgh Press. pp. 215–227.

d'Apollonia, Sylvia, & Abrami, Philip C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*(11), 1198–1208. doi:10.1037/0003-066X.52.11.1198

Darby, Jenny A. (2008). Course evaluations: a tendency to respond 'favourably' on scales? *Quality Assurance in Education, 16*(1), 7–18. doi:10.1108/09684880810848387

Dayton, Amy E. (2015a). Assessing teaching: A changing landscape. In Amy E. Dayton (Ed.), *Assessing the teaching of writing* (pp. 1–12). Logan, UT: Utah State University Press. doi:10.7330/9780874219661.c003

Dayton, Amy E. (2015b). Making sense (and making use) of student evaluations. In Amy E. Dayton (Ed.), *Assessing the teaching of writing* (pp. 1–12). Logan, UT: Utah State University Press. doi:10.7330/9780874219661.c001

Dewar, Jacqueline M. (2011). Helping stakeholders understand the limitations of SRT data: Are we doing enough? *Journal of Faculty Development, 25*(3), 40–44.

Emery, Charles R., Kramer, Tracy R., & Tian, Robert G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, *11*(1), 37–46. doi:10.1108/09684880310462074

Feinstein, Alvan R., & Cicchetti, Domenic V. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43,* 543–549. doi:10.1016/0895-4356(90)90158-L

Franklin, Jennifer L., & Theall, Michael. (1990). Communicating student ratings to decision makers: Design for good practice. *New Directions for Teaching and Learning*, *1990*(43), 75–93. doi:10.1002/tl.37219904308

Gorgias. (1991). Encomium of Helen. In G. A. Kennedy (Trans.) *On rhetoric: A theory of discourse* (pp. 283–288). Oxford: Oxford University Press.

Gravestock, Pamela, & Gregor-Greenleaf, Emily. (2008). *Student course evaluations: Research, models and trends*. Toronto: Higher Education Quality Council

of Ontario. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.627.559&rep=rep1&type=pdf

Guggenmoos-Holzmann, Irene. (1993). How reliable are chance-corrected measures of agreement? *Statistics in Medicine, 12*, 2191–2205. doi:10.1002/sim.4780122305

Harpe, Spencer E. (2015). How to analyse Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning, 7*(6), 836–850. doi:10.1016/j.cptl.2015.08.001

Harrison, Paul D., Douglas, Deanna K., & Burdsal, Charles A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, *45*(3), 311–323. doi:10.1023/B:RIHE.0000019592.78752.da

Li, Dan, & Benton, Stephen L. (2017). The effects of instructor gender and discipline group on student ratings of instruction. *The IDEA Research Report #10*. Retrieved from http://www.ideaedu.org/Portals/0/Uploads/Documents/Research%20Reports/Research_Report_10.pdf

Linse, Angela R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*. *54*, 94–106. doi:10.1016/j.stueduc.2016.12.004

Marsh, Herbert W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In Raymond P. Perry & John C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/1-4020-5742-3_9

McCullough, B.D., & Radson, Darrell. (2011). Analysing student evaluations of teaching: comparing means and proportions. *Evaluation & Research in Education, 24*(3), 183–202. doi:10.1080/09500790.2011.603411

Mitry, Darryl J., & Smith, David E. (2014). Student evaluations of faculty members: A call for analytical prudence. *Journal on Excellence in College Teaching, 25*(2), 56–67.

Moore, Cindy. (2015). Administrative priorities and the case for multiple methods. In Amy E. Dayton (Ed.), *Assessing the teaching of writing* (pp. 133–151). Logan, UT: Utah State University Press. doi:10.7330/9780874219661.c009.

Neumann, Ruth. (2000). Communicating student evaluation of teaching results: Rating interpretation guides (RIGs). *Assessment & Evaluation in Higher Education, 25*(2), 121–134. doi:10.1080/02602930050031289

Palmer, Stuart. (2012). Student evaluation of teaching: Keeping in touch with reality. *Quality in Higher Education*, *18*(3), 297–311. doi:10.1080/13538322.2012.730336

Penny, Angela R. & Coe, Robert. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Education Research, 74*(2), 215–235. doi:10.3102/00346543074002215

Samuels, Bob. (2018). Contingent faculty and academic freedom in the age of Trump: Organizing the disenfranchised is the key to success. *Forum: Issues*

About Part-Time and Contingent Labor, 21(2), A21–A24. Retrieved from http://cccc.ncte.org/cccc/forum/issues

Schweitzer, Leah. (2009). Accommodating the consumer-student. *Composition Forum, 20*. Retrieved from http://compositionforum.com/issue/20/accommodating-consumer-student.php

Spooren, Pieter, Brockx, Bert, & Mortelmans, Dimitri. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598–642. doi:10.3102/0034654313496870

Sullivan, Gail M., & Artino, Anthony R. (2013). Analysing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education, 5*(4), 541–542. doi:10.4300/JGME-5-4-18

Thorne, Gaylord L. (1980). Student ratings of instructors: from scores to administrative decisions. *The Journal of Higher Education*, *51*(2), 207. doi:10.2307/1981375

Alexis Teagarden is the director of first-year English and an assistant professor of English at the University of Massachusetts Dartmouth. Her current research projects focus on information literacy/research skills, source synthesis, and faculty development and evaluation.

Michael Carlozzi was a first-year writing instructor at the University of Massachusetts Dartmouth. His research interests include assessment, library science, and the perfect game of cribbage.