Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 1, February 2022 ISSN 1531-7714

Gauging Teaching Performance: Observational Sampling Opportunity, Reliability, and the Manifestation of True-Response Data

Jeffrey N. Howard, *Northern State University*

The Student Evaluation of Teaching (SET) instrument provides insight for instructors and administrators alike, often touting high response-rates to endorse their validity and reliability. However, response-rate alone omits consideration for *adequate quantity* of *'observational sampling opportunity' (OSO) data points* (e.g., high student attendance). The current paper endorses that quantity of OSO data points is critical to validity/reliability of longitudinal SET paradigms. It is reasoned ethically-challenged to rely on SET via basic surface-measures such as simple 'response-rate', when specific higher-quality data reflecting adequate quantity of OSO data points, can be filtered for from the same dataset. In addition, ethical concerns regarding the gauging of teaching performance via quantitative data analyses applied to inappropriate categorical/nominal response data, is also discussed.

The Student Evaluation of Teaching (SET) mechanism for evaluating instructor performance has been a mainstay in higher education for decades. The SET tool is well known for its inclusion as a decision-facilitating apparatus in everything from the quest for tenure, to simple base-level assessment of non-tenured faculty and instructors. With respect to the decision-facilitating nature of the SET, it is often the case that SET response reports from students providing the data are 'rolled up', and this composite report is then utilized as the decision-making tool by those individuals and/or committees in supervisory or evaluative positions. However, the manner in which the SET data reports are often compiled, analyzed, and evaluated leaves much to be desired—both on a figurative level, and a statistical validity level. One can readily argue that heuristics such as 'representativeness' lead us to make decisions based on small sample sizes, and that people often fall victim to small-sample fallacy out of assumption that such samples are representative of their parent population (Matlin & Farmer, 2016). Thus, as decision-makers, we may naturally overlook the Observational Sampling Opportunity (OSO) data points of a report reflecting an observational sequence, only to emphasize the number of reports we have accumulated across said observational sequence as validly supporting what was observed. In short, we believe that merely having a large number of reports on an observational sequence is itself somehow representative of *the number of potential OSO data points these reports are capable of being comprised of*—when in fact such a belief is false.

Part of the reason for such false belief is that quite often there are no individual class-by-class data reports for student SET observations over a semester. And albeit, there exists no separate data report for each class, the SET design is presumed to access the collective experiential memory of the student across a semester, which culminates in an end-of-semester recall-report based on this student collective memory data. However, there does exist a metric that allows for maximizing these aforementioned potential OSO data points contained within individual SET reports—and that metric is 'student attendance'—the same metric that has been empirically shown to enhance student examination performance (Purcell, 2007); the same metric that is

consistent with theories of learning that stress the critical impact of repetitive skills practice, and repetitive and extensive exposure to information (Credé, Roch, & Kieszczynka,2010).

Reliability and Data Resolution: Data Point Quantity vs Data Point Quality

One of the most inherent and pervasive issues within the observational data collection of the teaching-evaluation paradigm, is the emphasis on the quantity of submitted data-reports, over the emphasis on the quantity of data contained within each of those reports—the OSO. Stressing the quantity of reports accumulated is to endorse a top-level population parameter that does not accentuate high-resolution characteristics of the underlying data. In essence, population parameters are 'emergent properties' of their population whereby reality is organized in levels, and properties that are novel arise from that which manifests at lower levels (Arocha, 2020). Given that it is the properties of individuals that give rise to psychological processes, to study data about aggregates, that foregoes emphasis on individuals, "does not qualify as psychology" (Arocha, 2020; Lamiell, 2018, p.491).

The following scenario demonstrates the importance of longitudinal OSO data point *quantity* in the quest to achieve aggregate observational data *quality*:

> Imagine an observational data-collection paradigm where 15 Major League Baseball (MLB) scouts from 15 different teams show up to a championship college game to watch a starting pitcher prospect from a certain team. Five of the scouts leave after watching the prospect pitch for 2 innings; seven more of the scouts leave after watching the prospect pitch 6 innings; two of the scouts leave after watching the prospect through 7 innings; the sole remaining scout watches the pitcher finish a complete 9-inning game for the win.

Suppose as a sport's scientist, you were provided with and had the chance to analyze the anonymous data from all of the scouting reports that were actually turned in, for a paper you wished to write and publish. First, consider this: which scout—based on longitudinal observation across time (e.g. number of innings)—would you think to be most qualified in judging the ability of the pitching prospect? Twelve of the scouts turned in their scouting reports—an 80% response rate—but *which* 12 scouts are the ones who turned in those reports that comprise the anonymous aggregate data set you possess? In an anonymous data aggregation scenario like this (or like the SET's), one does not know. Would one simply report the 'terrific' 80% return rate of scouting reports when submitting their manuscript for publication, or would one be concerned with the *underlying quantity of observational sampling opportunity data points* reflected by each of those 12 reports (e.g. innings spent observing individual pitches thrown)?

It is highly unlikely that one's paper would be recommended for publication based on simple scouting report return-rate that was highly suspect regarding its underlying observational data point quantity and composition. A competent PhD-level statistician would be highly suspect to make any such recommendation in this scenario if they knew of the true underlying data composition and its compromised aggregation process. Yet, this scenario is exactly what unfolds within many anonymous SET student data reporting, aggregation, and reporting sequences. In short, the data are questionable in statistical validity in that it is errantly represented by a 'surface-level' measurement (simple survey response-rate) that obscures the true quantity and integrity of the individual OSO data points that it is comprised of.

Unsurprisingly, such anonymous aggregated compromised data—data lacking in longitudinal observational data-point *quantity*—is exactly what SET reports are constructed from; reports that administrators, tenure committees, and instructors are consulting with respect to quality of instructor performance.

Reliability and Sample Size: Quantity of Observational Sampling Opportunity

A solution-oriented perspective to the previously introduced issue of low OSO manifestation within lower-level subset data, is rather apparent: student attendance. Given that the very nature of generating an OSO is to attend class, placing constraints on survey participation relative to class attendance, could provide a needed boost in OSO quantity and quality at a lower sub-aggregate level in the data. The status quo of pooling all student SET reports and reporting surface-level aggregate results like 'response rate' is inaccurate at best and should be remedied by greater quality controls—controls that emphasize 'maximum opportunities to observe' for those reporting SET response data. To

resolve such an issue, the suggestion is put forth that the SET completion opportunity must be 'earned' by students via high attendance—attendance in each class represents a 'data observation opportunity'—such opportunities are that which constitute 'quantity' of longitudinal observational data-points. Or perhaps all students could complete the SET, but only those reports from students meeting the 'attendance' criteria are selected and processed for data aggregation.

For example, only those students meeting criteria of having attended 80% of the classes or more would be provided the opportunity to complete the SET instrument. If an instructor is truly of poor quality, then a greater quantity of OSO data points (e.g. days attending class) on the part of student-raters would have far greater validity in demonstrating such shortcoming in performance. Likewise, if an instructor is of high quality, the data would have elevated validity in confirming this as well.

Allowing students who may have only attended 20-25% or so of classes to complete a SET is not ethical in that it is easily mathematically demonstrable that such lack of observational data point quantity could not muster a sample size reflective of high validity. Anecdotally, it has been rumored that some students who have enrolled in a course, attended the course only once or twice, and having decided they did not like the instructor, dropped the class—only to be sent an electronic link via university email soliciting them to complete the SET. If one reverses the perspective, the ethics of such a situation gains clarity: imagine a student turns in a research paper 20 pages in length, and the instructor reads the first 3 pages, and subsequently assigns the student a failing grade. This meager 'exposure' to the student's writing ability would be in equal proportion to a student attending 6 of 40 possible class meetings, dropping the class, and giving the instructor a 'failing evaluation'. Yet, virtually no teacher would agree with the ethics of giving a student a failing grade on a research paper, after having read only 15% of said paper.

The bottom line here is that a reporting mechanism other than the SET should be considered for instances when the student has dropped the class due to an early 'falling-out' with the instructor. Such feedback design would avoid pooling low-resolution OSO data with high-resolution OSO data.

Most would agree that if students were collecting mere frequency-count data on how many times an instructor took a drink from their water bottle during class, then pooling such data between low-classroom attendance students and high-classroom attendance students is acceptable and of little concern. However, *qualitative observational data* derived from subjective-judgments reflective of the characteristics of the instructor and an environment controlled by said instructor, should not be pooled between low-attendance and high attendance students—and it is ethically-challenged to do so. Even more so, it would be reasoned unethical to use such diluted pooled data to make employee quality/advancement decisions.

Perhaps a counter-argument from some may be the hierarchical nature of the instructor-student relationship and its influence over students—the idea being that since instructors have power and influence over students, all students should in some way be automatically compensated for this power imbalance. However, this idea is based on pure 'speculation'—speculation that instructor influence and power over students will always manifest as 'fear of retribution' within students, causing students to give instructors good ratings when those instructors truly do not deserve them. This is noticeably false based on the abundance of negative qualitative comments about instructors and instructor performance, which are at times immediately followed on the SET commentary report by contradicting and glowing qualitative comments regarding the instructor's stellar performance and overarching fairness. In addition, there exist adequate administrative means for addressing student complaints regarding any instructor retribution that may have indeed occurred—some of which can even provide student anonymity at lower-levels of the complaint process, thus neutralizing any potential 'retribution' that could occur.

Garnering Data Integrity: Earning the SET right

Student attendance is known to be a component of the ubiquitously-worshiped highly-desirable outcome known as 'student engagement' (Beran & Violato, 2009), and without a doubt, making attendance mandatory could better assure high-quantity of OSO data—research has demonstrated such a connection (Schlenker, & Coles McKinnon, 1994). But such a requirement would not prevent an unmotivated and/or vindictive student(s) from attending the class only 10-

15% of the time, submitting a SET report that is negative, and having that negative low-quality data dilute other high-quantity OSO data. However, filtering for and considering only those SET reports that meet OSO quantity criteria (say for example, 80% + attendance) would indeed avoid such dilution, thus preserving data integrity and bolstering validity of the instrument. In fact, according to Davidovitch & Soen (2006) in a study of 9,636 questionnaires reflecting student rating content across 634 differing courses, it was found that getting students to attend class more could reliably and factually translate into higher ratings for teachers. Thus, when it comes to facilitating equitable instructor ratings, a suggested sample value of an 80% attendance rate threshold, as per the aforementioned research support, would be a reasonable estimate.

For example, if each class period were counted as an observational period, within a 16-week semester there would be roughly 42 Monday-Wednesday-Friday class periods for observation of the instructor—16 weeks minus finals-week (no classes); minus roughly 3 class periods (one week) for federal holidays/breaks etc. This comes to 14 weeks of observational periods, times 3 periods per week, which equals ~ 42 observational periods. With a class of 30 students there would be 30 x 42 = 1260 independent observational periods possible by all students combined. Therefore, hypothetically using 80% attendance rate minimum criteria as an example, for any one student, there would be .80 x 42 = 33.6 (34) observational periods (or roughly 8 days of allotted absence per student).

Students would thus need to have eight or fewer documented absences in order to have their SET data included in the final dataset aggregation. Student attendance could be reported by instructors to administrators who are responsible for sending out solicitations for completion of the SET's, and only those meeting attendance criteria would be sent the link to solicit SET completion; or all students could be sent the link to solicit SET completion, and filtering for qualified reports based on attendance could be made after-the-fact if attendance rolls could be matched against SET reports, and SET reports at this stage are not anonymous.

As for the actual physical process of taking attendance, instructors could both 'call roll' and pass around a 'sign-in' sheet so as to have a cross-corroborating system of attendance should there be any question as to instructor reporting of the numbers. At the extreme, student ID cards could be scanned by handheld scanners passed around the class or located near doors—this data would be automatically uploaded to the system, and would be inclusive of information that would be impossible for faculty to forge/manipulate, thus validating actual attendance with a data-encryption level of accuracy and security.

Faculty-supervising administrators may not agree with limiting data by only allowing those who meet some predefined attendance criteria to fill out SET's—but although the quantity of data via response rate may be reduced, the quality of the data that remains has higher integrity and greater validity than any 'diluted' form of such data. Administrators may find this to be an attractive compromise.

At issue is also the fact that SET instruments may not be completed by all students immediately after the last day of class—some might be completed several class periods ahead of the end of the semester if links to the instrument were sent out two weeks in advance. This could diminish the 42 observation periods down to 35 or so, which clearly cuts into the number of OSO data points available.

Leaving completion of SET reports up to students despite their attendance status in the class could be a challenge as well. Students with high attendance still may not complete the SET unless compelled to do so. A solution may be to require all students to complete the SET instrument before their final grades could be entered into the system—thus it becomes a mandatory required responsibility of all students regardless of their performance or attendance in the class.

Some might put forth the argument that by doing this, one is compelling participants to participate in research, and thus they may question the ethics of such compulsion. However, students are the *researchers* in the SET paradigm, and not the *participants*—this is clear when one considers who the data are actually about, and who is completing the 'research reports' (SET's) (Howard, 2021). 45-CFR-46 is clear by its fundamental definition about requiring that research protections always *lean toward* research subjects, and *away from* researchers, therefore the predominance of protections in the SET paradigm should favor instructors and not students (Howard, 2021).

Under the proposed data-filtering change, students would have the option to accumulate high-attendance, and complete the SET, despite their displeasure with the course and instructor—a known contributor to absenteeism (Treischl & Wolbring, 2017)—the idea being that if a student is willing to go to a class that is truly a poorly taught class, by a poor instructor, then the SET evaluation carries with it high validity. In short, to truly lodge a 'valid' complaint via SET regarding instructor performance, there should be a substantial investment on the part of the student—adequate OSO quantity—quantity that meets specific sample size criteria. This is balanced out by the fact that there may be a substantial price to pay on the part of the instructor for poor evaluation on the SET: employment termination. In allowing a student to lodge a SET complaint when dropping a class early, any investment on the part of the student is circumvented and their SET report should be highly suspect, if not entirely invalid.

Additionally, one has to consider those who are already highly-challenged in class attendance to begin with if one were to utilize a design with attendance as a determinant in the evaluative process. Students who are disabled, or for other valid reasons are unable to attend and/or possess a disability that limits their ability to make 'observation' in some way—such as visual impairment—would clearly require that a valid parallel means to complete observations be designed when and where possible. Also, some courses would need their own customization implemented so as to provide a relative system-equivalent attendance-measurement outcome to equate them with in-class attendance patterns of the regular classroom—such as courses in-the-field or in-the-lab that may not be conducive to highly consistent in-class attendance.

One final factor that should not be overlooked is the lack of student training in constructive feedback and evaluative processes. Even with high/perfect attendance, students with lack of training on observational research and constructive feedback techniques may not provide the highest quality assessment data when it comes to SET's. Hartenian (2016) has suggested that training applied within new student orientation on a goal-oriented system of providing instructor evaluation could be effective in giving further credence to student evaluation of instructors. With training early on in their academic career, students could be taught to assess an instructor with respect to their performance and skills within the context of the instructional environment and the topics at hand, and to ignore/neutralize personal characteristics, personal biases, or other superficial irrelevant information that may influence instructor ratings.

Threats to Validity: Multi-Tiered Data-Aggregation

In pooling the data from low-quantity and high-quantity behavior observation datasets, one has watered-down that which can achieve highest validity within the instructor evaluation. This is the standard operating procedure of many SET paradigms, and it can result in sketchy performance profiles for those who truly perform at high levels. High-quantity observational data represents student SETs whereby there exists a high number of 'opportunities to observe' the instructor in the classroom environment. An aforementioned suggested criterion was 80% or higher attendance level during a semester. However, despite the fact that one could indeed filter for student SET reports that represent 80% attendance or higher, actual SET data in aggregated reports wantonly pool low-attendance and high-attendance researcher (student) SET reports. There is little reason as to why the ethics of such pooling of low-quality and high-quality qualitative response data should be left unchallenged.

The field of Applied Behavior Analysis (ABA) conveniently highlights how data-aggregation affects lower-level data characteristics such as high-quantity vs low-quantity of OSO data:

> Throughout the history of behavior analysis, however, the problem of perceiving differences in the face of variability has meant that statistical aggregation has subtly crept in. Any form of aggregation helps to smooth the data thus making patterns easier to see, but it also can hide trial-by-trial variability and sample size information. Variability, sample size, and other data characteristics should be informing a scientist's inferences but cannot do so when they are not faithfully represented (Young, 2018).

Bringing the aforementioned quote to life, Godwin et al. (2016) provide a good example of adjusting for sample size within data analyses so as to make consideration for accurate statistical power—which in

this case, without adjustment, would be overinflated and not highly representative of the dataset:

> Therefore, a total of 2,402 student-session pairs were observed. A student session pair refers to a specific student observed by a coder within a specific session. However, treating the children within each session as a different set of students artificially inflates statistical power. In order to mitigate this concern, a more conservative alpha level was used in the analyses reported below. Specifically, the alpha level was adjusted to 0.0083 (the commonly accepted alpha level of 0.05 was divided by 6, the total number of observations, in order to more closely approximate the true size of the sample) (Godwin, et. al, 2016, p.131).

Such adjustments can be easily made on data prior to analyses, rather than within analyses procedures themselves—such as filtering for 'complete' datasets or datasets that are complete up to or beyond some predefined criteria (e.g. 80% attendance or higher), so as to avoid intentionally diluting a dataset by merging low-quantity student OSO data with high-quantity student OSO data.

It also serves to point out that the concept of compiling student reports reflecting student observational experience longitudinally across a semester, is in essence an inter-rater agreement exercise—and inter-rater agreement for observational data has known specific data-analysis requirements that must be fulfilled (see Hallgren, 2012; Walter, Eliasziw & Donner, 1998). Thus, departure from criteria that assure integrity at all levels of resolution within SET data, could clearly yield compromised inter-rater agreement.

In addition to data resolution integrity issues, according to Young (2018) there exists a dearth of history regarding the ability of people to make proper inference via correct integration of data features; in particular: components such as intrasubject variability; inter-subject variability; sample size; distribution properties of response variables (e.g. normality, skewness); relationship nature/magnitude; variables that moderate, and data dependencies—all of which pose critical to correct inference. It's quite clear there would be a similar 'dearth of experience' on the part of amateur student observers, with respect to such data feature components and making proper inference about behavioral observations of educators within a classroom. Pollett, Stulp, Henzi, & Barrett (2015) refer to such improper inferences as the 'Ecological Fallacy' (EF) whereby relationships at the level of the individual can be of a different magnitude than those at differing levels of aggregation. Even more telling is a special case of EF—known as 'Simpson Paradox' (Simpson, 1951)—which has demonstrated that a direction of a relationship within a number of individual groups, can be reversed when applying the same analysis at the population level. It is also interesting to point out, with respect to student observation and inference, that the field of ABA—known for its ubiquitous observational technique of measuring 'off-task' and 'on-task' child behaviors in the classroom—serves as an excellent parallel-example for the off-task wandering/rambling professor; a criticism that may well be the one of the most commonly mentioned teacher shortcomings in SET reports.

When it comes to solutions for the pooling/aggregation of multi-tiered data, the fields of epidemiology and sociology, via mixed models, random coefficient models, random effect models, hierarchical models, and nested models, have also shown efficacy in addressing data-aggregation concerns. In addition, Pollet et al. (2015) indicate that perhaps one of the most important tools in facilitating such results is the ability of modern statistical software packages to address these types of multilevel analyses more effectively. Thus, it seems quite reasonable that one of the best combinatorial solutions to any unwanted influence on the data that might arise from data aggregation techniques, is better data-collection design principles applied in-tandem with modern software analyses strategies.

Threats to Validity: Bias in Behavior Observation Paradigms

The process of behavioral observation leading up to SET reporting is fraught with bias potential in both the cognitive realm, and within the realm of statistical sampling. Human nature is such that cognitive processing and its bent towards processing efficiency, is an ideal proving ground for such biases. For example, humans are not highly proficient with respect to making good social-perceiver inference from samples; are inattentive to the size of samples, and fail to recognize that large samples are much better than small samples in estimating characteristics of a population (Fiske & Taylor, 2017). As it relates to prediction of future behavior, (Fiske & Taylor, 2017) state the following: "...people will often overgeneralize from a small

unrepresentative sample. For example, on witnessing a single instance of another person's behavior, the social perceiver will often make confident predictions about that person's behavior in the future. [Fiske & Taylor, p.223]." This 'small-sample fallacy' as it is also known, is directly connected to the earlier mentioned concept of 'representativeness' and is often manifest within social situations (Matlin & Farmer, 2016).

Also demonstrating interjection of potential bias are judgements based on delayed memory recall, as opposed to recall rendered with high immediacy. According to Hastie & Park (1986), judgment-outcomes have been shown to be correlated with memory when judgment is rendered from memory of an event—but are less correlated with memory when those judgments are rendered as immediate (on-line) judgments of the occurring event—thus demonstrating that the fallible nature of human memory can be mitigated by more consistent and contiguous contact between those rendering judgment, and those being judged.

The role and importance of the underlying sample data feeding an SET instrument cannot be underestimated. With inadequate sample data available, the idea that one can accurately draw inference is highly challenged. One such challenge is that of 'extreme examples' occurring within the sample data (Fiske & Taylor, 2017) With respect to the SET and opportunities to observe within the classroom, such extreme examples can have a greater effect on those who have small data samples (e.g. higher absenteeism). With poor attendance, students have less instructor behavioral data to draw from, and thus an in-class instructor mistake during lecture, a poor instructor lecture example, or a poor explanation of a learning concept in a lecture can be magnified and become an extreme (uncharacteristic of the instructor) example for those students. This could lead to inaccurate estimation as to how frequently such extreme examples occur within the range of instructor behaviors—an unfortunate occurrence given that estimation of behavioral frequency is critical with respect to using behavioral sample information to render judgements (Fiske & Taylor, 2017).

Attribution theory further demonstrates the problematic nature of missing data when making behavioral construal about an individual—such as a professor—regarding their personal nature and abilities. Kelley's (1973) 'covariation model' of attribution theory indicates that multiple behaviors are examined at different times and situations using three fundamental information components—consensus, distinctiveness, and consistency—to arrive at a decision as to whether behavior is due to the person themselves (internal attribution) or due to some other person, situation, and/or circumstances in-place at the time of the behavior (external attribution) (Aronson et al., 2015). Perhaps most important is the fact that extensive student absenteeism allows for far less examination of multiple behaviors across time (low sample size). Thus, in the case of high absenteeism, the default of 'internal attribution' is likely dominant when making decisions about the nature of the professor. Additionally, people use a two-stage attribution process when construing causation (Gilbert, 1991)—the first stage is to make the internal attribution; the second is to make adjustment to the first stage by considering situational factors influencing the person. However, becoming inattentive or losing focus in some way can cause omission of stage two, thus allowing extreme internal attribution to prevail in isolation (Aronson et al., 2015). One need exert little effort to see that persistent absenteeism from class may well be the pinnacle of such inattention and loss of focus.

There are indeed instances in the research where mention of individual underlying OSO data points does surface, but the concept of 'class attendance' that provides these individual OSO data points is seemingly evaded. Benton & Ryalls (2016), researchers for a well-known Student Rating of Instruction (SRI) instrument publisher, provide a good example:

> Because well-constructed SRI present multiple information from individuals (students) within a class and are collected across multiple occasions, one can make the case that students provide the most reliable source of feedback about teaching (Marsh, 2007). In contrast, class observations performed by an administrator or a peer—be they trained or untrained evaluators—typically represent only one observation on one occasion. In this case we do not know what the consistency/reliability is of their ratings. For reliability, trust SRI. (Benton & Ryalls, 2016, p.3).

Here the authors clearly acknowledge that it is individual student observation across 'multiple occasions' (high quantity of OSO) within student rating of instruction that provides the greatest evidence of reliability for the SET—yet in their paper they

demonstrate a persistent focus on sample size as it relates to mere 'response rate', which is mentioned 13 times; the paper does not address nor even mention the concept of 'student attendance/absenteeism'.

Going further, these same authors indicate the following regarding 'reliability' and the number of student raters:

> Reliability, on the other hand, is related to sample size, or the number of student raters. If 50 students out of a class of 100 responded to a survey, their ratings would be more statistically reliable than if 19 students out of a class of 20 responded even though the 19 responders would be more representative (Benton & Ryalls, 2016, p.3).

The authors point out that reliability is related to sample size as derived from the number of student raters who turn in a completed survey. However, they overlook the critical fact that statistical concepts—such as sample size—that apply at the level of student response rate, also apply at higher resolution levels deeper within the data. Each student rater who responds with a completed ratings survey has submitted a survey that is also comprised of an underlying sample size of 'X' *observational sampling opportunities—or OSO's*. Thus, a large sample size of completed student rater surveys, each of which is comprised of small inadequate samples of individual observational data, would be compromised at its most basic level—the level of 'individual opportunity to observe'—which is also known as 'class attendance'. Here, in the authors' example, if the 19 students actually went to class 90% of the time, and the 50 students only went to class 30% of the time, with a greater proportion of individual underlying data observation opportunities being accounted for by the 19 students, there could well be a 'cancellation factor' that might render the reliability coefficient of the two disparate sample sizes as truly equivalent. A high-impact, high-relevance point regarding reliability is provided by Wilhelm, Rouse, & Jones (2018, p.2), via conjugation of insight from van der Lans et al. (2016), and Krippendorff (2016):

> Even when a generalizability study has been conducted to recommend the number of raters, the number of observations, and the level of training required of raters, the use of a validated observational system does not ensure that the data produced will be reliable (van der Lans, et al., 2016). The critical final piece is ensuring that the rating process has not produced irrelevant variation. Demonstrating agreement between replications by different raters "allows us to infer the extent to which data can be considered as reliable surrogates for phenomena of analytical interest." (Krippendorff, 2016, p.139)

This begs the question: How can said 'reliable surrogacy' occur across replications via different raters displaying extreme and varying degrees of exposure to the 'phenomena of analytical interest'?

When it comes to cognitive biases, there is a plethora which might be alluded to when it comes to mechanisms and instruments of observation and evaluation (see Tversky & Kahneman, 1974). However, one type of bias that may well be at the forefront of SET ratings influence factors is the *self-serving bias*. The self-serving bias is that whereby one creates attributions about causation of events, outcomes, or action related to the self—attributing positive outcomes to the self and one's disposition, and negative outcomes to external and situational factors (Forsyth, 2008). Such bias may tend to arise as a protection to the ego within the midst of, or awareness of, poor performance or other occurrence which may threaten the 'self' (Forsyth, 2008). Albeit the term 'self-serving' promotes the concept of 'self', such bias is indeed capable of extending outward away from behavioral explanation of oneself, to the perception of friends, partners, or even groups that one belongs to (Fiske & Taylor, 2017). Thus, one could clearly attribute one's overall success in a class due to oneself and one's internal disposition, and similarly, attribute one's failure(s) in the class externally—to one's instructor—an attribution that would clearly lead to negative SET results.

Although the self-serving bias likely carries a great deal of influence in the SET paradigm due to anonymity that students have in rendering evaluations, Goos & Salomons (2017) put forth a strong argument for 'selection-bias'—a bias that provides opportunity for the self-serving bias to occur—as a factor exerting influence over precise SET reflection. The concern is that without random selection of students for SET participation, positive selection bias reflects SET scores that are actually lower. Such positive selection bias is driven by characteristics inherent in student subsets such as observable variables of grade, gender, and course size Goos & Salomons (2017). One can use the example of rating service satisfaction at a supermarket as a parallel: when we receive what we intended to get, at the price we

find reasonable, there appears little need to complain or expend extra effort to rate what we expected to receive physically and financially as a result of the transaction. Similarly, when students receive a grade they tried to get (or are happy with), with the effort they felt was reasonable, they may not respond to a SET invite email; those performing poorly however, may feel quite the opposite. The point that Goos & Salamons (2017) make is valid—*both* of these performers should be solicited randomly to render SET report data, so as to attempt to gain more equal representation of both types of respondents within the SET database. Likewise, if variables such as gender vary in their natural tendency to respond to the SET opportunity, then a random sampling across genders for the SET invite would also be appropriate. In fact, Valencia (2020) investigated the variables of both gender and 'acquiescence' (a tendency to render favorable response options across items) in-tandem, with results indicating that acquiescence tends to inflate item response values, with differences in teaching quality being reduced between female and male instructors. Thus, the revelation of gender differences is unnaturally compressed by such 'acquiescence effect'.

Similarly, there can also be a 'halo-effect' presence that influences outcomes via cross-contamination of questions. However, a halo-effect may not be as impactful as other effects, such as gender-acquiescence interaction. Cannon & Cipriani (2021) found that such halo-effect are present in SET instruments, but also point out that the informative integrity of the SET is relatively preserved, and as per their results they do not offer up recommendation that restructuring of SET instruments such result due to the halo-effect.

Naturally, it behooves the researcher to make in-depth consideration regarding extraneous variability due to variable interactions and/or variables in isolation, when it comes to instrument design—not the least of which should be the 'fundamental resolution quality' of any/all rendered dataset items, as purported by the base-purpose of this paper.

Statistical Validity: Results That May Not Truly Manifest in the Data

Perhaps most important is the tenuous idea that performance evaluations are solidly derived from SET data gleaned from statistically sound instrument design and accompanied by proper application of measurement-scale analyses. In many cases, this departs from reality being that the impropriety of interpreting means and standard deviations from categorical/nominal data, appears to be alive and well. Stark and Freshtat (2014) point to this issue reminding that student evaluation of teaching is often represented by errant data analyses applied to numerical values that actually represent categorical/nominal data as 'labels.' Such categorical/nominal scale data do not demonstrate equivalent distances between their membership categories (e.g., the scale distance between 'blonde', 'brunette', and 'redhead' may not be perceived as equivalent). The following example illustrates a form of this data-analysis misapplication—imagine an experiment with the following methodology and results:

> [1] A computer presentation of visual stimuli consists of 7 on-screen colored/numbered boxes where 1=red, 2=orange, 3=yellow, 4=green, 5=blue, 6=indigo, 7=violet;
>
> [2] 20 participants are asked to select a colored box for each randomly presented tone they hear on 10 trials—200 total trials (100 per group);
>
> [3] Each corresponding color number is recorded into the datafile to represent the color that was selected on a particular trial;
>
> [4] 10 of the participants always select 1 (red) when they hear a tone;
>
> [5] 10 of the participants always select 7 (violet) when they hear a tone;
>
> [6] The 10 participants who chose 1 (red) on all 10 trials would amass a total score of 100;
>
> [7] The 10 participants who always chose 7 (violet) on all 10 trials would amass a total score of 700.

The results of such methodology, under the application of data analyses to extract the 'average' color-choice across trials would be as follows: the grand total of all trial values would be 100 + 700 = 800; this grand total of 800 divided by 200 trials = 4—thus the 'average color choice' across all 200 trials would equal '4', which coincides with the color 'green'—albeit nobody selected 'green' during the entire 200-trial experiment!

Such adding-up and averaging of categorical /nominal variables represented by numerical label values, would never occur in a sporting event television broadcast using numbers on player jersey's (labels) because even the layperson knows that such statistical

calculation 'would not make sense'. The bottom-line is that there exist myriad teaching evaluation instruments that assign numerical values to categorical scale responses such as 1=hardly ever, 2=sometimes, 3=occasionally, 4=frequently, and 5=almost always, and then use the 'average' of those categorical/nominal label values to quantitatively assess performance—assessment that could errantly reflect values that may not even exist within the master dataset of responses generated by those making the ratings.

Going one step further, it is known that the strategy of combining categorical data via a process of 'collapsing down' multi-item Likert-scale categories into fewer categories (e.g. agree vs disagree) may also be applied—a process that reduces the resolution of the instrument's OSO data-points, thus failing to preserve the true relationship within question items and responses (Palmer, 2012).

One recommended solution to the issue of applying quantitative analyses to categorical/nominal variables, would be to use a continuous variable measurement approach for participant scoring of each item. For example, rather than using a Likert-scale with items 1=hardly ever, 2=sometimes, 3=occasionally, 4=frequently, and 5=almost always, one could implement a line-based continuous design such as:

hardly-ever _____ almost-always

By instructing participants to make a mark on the line where their response would be located, all possible values are encompassed, thus choice perception is represented as contiguous ratio-scale data—a mark at the very left end could be '0'; a mark at the far-right end could be '5'; resolution of values could be to two decimal places (e.g. 3.38, 2.76, 1.73). Categories of 'hardly ever', 'sometimes', 'occasionally' etc. could then be defined later within statistical analyses software by recoding variables that represent 'ranges' within the rendered ratio-scale data. Given that choices rendered would have been based on participant perception of a non-discrete continuous-scale of measurement, grouping them into recoded categorial variables by value-range after-the-fact (if one desires to do so), would be more appropriate.

However, in implementing such a data-collection design, one would need to physically measure the location of each mark on the 2-inch line with respect to the start of the line—a time consuming manual task that is best done via a web-interface type of 'slider' widget design that records this distance automatically when choice is rendered. Using this method, participant perception of choice is not limited to low-resolution discrete values, but rather participants perceive choice options as representing all possible values encapsulated by the continuum. And as a bonus, such design could actually include a numerical on-screen readout that changes as participants slide the widget along the measurement continuum, thus providing a perceptual mechanism demonstrating continuous-data high-resolution choice to two decimal places. In this manner, via this design, participant 'perception' of measurement is highly quantitative in nature, and thus quantitative analyses reported on such data would match participant perception at the point of rendering choice data.

By no means are the mechanics of and the perception of rendering data choice or OSO quantity the biggest sources of variation in the behavioral-ratings paradigm. Other sources of variation that contribute to error have well-documented support (see Borkan, 2017) such as teacher gender, student expected grade, characteristics of the rater (rater variability), characteristics of the teacher (weight, vocal-pitch, etc.), teacher mood, rater mood, etc.

The subjective nature of the rater can also influence rating-scale outcomes (An, Curby, & Brock, 2019) as can the phenomena of increased teacher-attentive reinforcement of student behaviors due to teacher a priori knowledge creating expectation effects that can influence the outcome—known as the 'Rosenthal Effect' (APA, 2021), and the converse—student-reinforcement of teacher behavior due to expectation effects on the part of students. The Rosenthal Effect in the form of student expectation effects may be even more common today than in the past due to websites like RateMyProfessors.com. Such websites provide past student commentary and ratings on professor performance, knowledge which might then influence future student expectation of teacher performance before a semester even begins. Similarly, teachers themselves—such as within the field of educational pedagogy—may be exposed to student performance bias ahead of time via 'dispositions' data-collection instruments on said students as required by teacher-education programs.

Conclusion

Granted the SET may only be one piece of a multi-faceted toolbox to assess instructor performance, and thus reliance on it and its potential impact in diluting the full range of instructor performance assessment data, may be mitigated across the full range of tools. However, such 'diversity-in-measurement' practices do not change the ethics of pooling low sample size and high sample size qualitative OSO data. It is curious to see research doctrine from an evaluation instrument publisher acknowledge the importance of 'multiple occasions of observation' (attending class) amongst student raters (Benton & Ryalls, 2016), only to have that research doctrine shift gears and leave the conversation about OSO data points behind in favor of 'response rate' sample-size endorsement. However, ignoring the importance of such underlying higher resolution data in favor of 'response rate' appears to be both a common theme in the research, as well as an anecdotal theme when one converses with other professionals about the subject. Even more curious is the fact that true administrative concern for high SET response-rate appears 'feigned' when one looks at delivery method, as there has been a large-scale shift to online administration despite evidence indicating in-class paper-based response-rate significantly surpasses that of online delivery (Capa-Aydin, 2016; Nulty, 2008; Ahmad, 2018). The big takeaway is that, when it comes to response-rate, one must avoid operating on any assumption that the pattern of responses for those not participating at all in instructor evaluation surveys, would somehow parallel that of those who do participate. Even more so, the assumption that low response rate is the fault of the professor, is baseless (Lawrence, 2018).

Ultimately, when it comes to observational data collection over time, it is of paramount importance to consider the various levels of resolution that may exist in the data—particularly when relying on sample size as an argument supportive of data reliability. Samples are quite often multi-tiered; it is good practice to require that sample-size integrity be present at all levels, especially if one wishes to make good argument supported by the data at aggregate levels.

With respect to an end-goal of recognizing threats to validity so as to catalyze change, it is suggested that a multidimensional strategy inclusive of: [1] acknowledging a need for student training as observers/researchers of human behavior as a component of a first-year student-success course; [2] incorporating a combination of experienced-observers and novel-observers in the class-observation report-generation process, and [3] a contribution of adequate sample-size student-generated OSO data at the highest possible resolution, would comprise the best rough-sketch plan to increase accuracy and scope of SET reports. It is also important to remember that research paradigms that are deemed 'exempt research' (e.g. SET's) are not automatically deemed 'research-that-cannot-harm'. Research protections must be extended to teachers within the SET environment so as to protect them from 'evaluative harm' via possible detrimental effects of unrepresentative data due to unfocused statistical aggregation, or as generated by raters with little exposure to those whom they are evaluating.

In the end, it is our lack of self-awareness as fallible beings who want to see only that which we want to see, at the resolution that we want to see it—students and teachers alike—that is our greatest obstacle when it comes to observing and reporting human behavior. As humans we are oft errantly focused on the end, rather than the means—failing to recognize that when it comes to a 'whole comprised of the sum of its parts', it is quite often the 'parts' that matter most.

References

Ahmad, T. (2018). Teaching evaluation and student response rate. *PSU Research Review*. https://www.emerald.com/insight/content/doi/10.1108/PRR-03-2018-0008/full/html

An, X., Curby, T. W., & Brock, L. L. (2019). Is the child really what's being rated? Sources of variance in teacher ratings of socioemotional skills. Journal of Psychoeducational Assessment, 37(7), 899-910.

APA Dictionary of Psychology: Rosenthal Effect. American Psychological Association (2021). https://dictionary.apa.org/rosenthal-effect

Arocha, J. F. (2020). Scientific realism and the issue of variability in behavior. Theory & Psychology, 0959354320935972.

Aronson, E., Wilson, T.D., Akert, R.M., & Sommers, S.R. (2015). Social Psychology. 9th Edition. Pearson. Boston, MA.

Benton, S. L., & Ryalls, K. R. (2016). Challenging Misconceptions about Student Ratings of Instruction. IDEA Paper# 58. *IDEA Center, Inc.*

Beran, T., & Violato, C. (2009). Student Ratings of Teaching Effectiveness: Student Engagement and Course Characteristics. *Canadian Journal of higher education*, *39*(1), 1-13.

Borkan, B. (2017). Exploring Variability Sources in Student Evaluation of Teaching via Many-Facet Rasch Model. *Eğitimde Ve Psikolojide Ölçme Ve Değerlendirme Dergisi*, *(Journal of Measurement and Evaluation in Education and Psychology) 8*(1), 15-33.

Cannon, E., & Cipriani, G. P. (2021). Quantifying halo effects in students' evaluation of teaching. *Assessment & Evaluation in Higher Education*, 1-14.

Capa-Aydin, Y. (2016). Student evaluation of instruction: comparison between in-class and online methods. *Assessment & Evaluation in Higher Education*, *41*(1), 112-126.

Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research*, *80*(2), 272-295.

Davidovitch, N., & Soen, D. (2006). Class Attendance and Students' Evaluation of Their College Instructors. *College Student Journal*, *40*(3).

Fiske, S. T., & Taylor, S. E. (2017). *Social cognition: From brains to culture*. 3rd Ed. Sage.

Forsyth, D. R. (2008). "Self-Serving Bias." International Encyclopedia of the Social Sciences. Edited by William A. Darity. 2nd ed. Vol. 7. Detroit: Macmillan.

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education*, *58*(4), 341-364.

Gilbert, D. T. (1991). How mental systems believe. *American psychologist*, *46*(2), 107.

Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A.V. (2016). Off-task behavior in elementary school children. *Learning and Instruction, 44*, 128-143.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23.

Hartenian, L. S. (2016). A Framework for Training Students as Evaluators of Instructor Performance. *Journal of Behavioral and Applied Management*, *3*(1).

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or online. *Psychological Review*, *93* (3), 258.

Howard, J.N. (2021). Anonymous Instructor Performance Evaluation: Tarasoff and 'Duty to Warn'. (manuscript in progress).

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107.

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology*, *12*(4), 139-144.

Lamiell, J. (2018). Rejoinder to Proctor and Xiong. *American Journal of Psychology*, *131*(4), 489-492.

Lawrence, J. W. (2018). Student evaluations of teaching are not valid. *American association of university professors*. May-June. https://www.aaup.org/comment/5555

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). Springer, Dordrecht.

Matlin, M.W., & Farmer, T.A. (2016). Cognition, 9th Ed. Hoboken NJ: John Wiley & Sons.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & evaluation in higher education*, *33*(3), 301-314.

Palmer, S. (2012). The performance of a student evaluation of teaching system. *Assessment & evaluation in higher education*, *37*(8), 975-985.

Pollet, T. V., Stulp, G., Henzi, S. P., & Barrett, L. (2015). Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. *American Journal of Primatology*, *77*(7), 727-740.

Purcell, P. (2007, September). Engineering student attendance at lectures: Effect on examination performance. In *International conference on engineering education–ICEE 2007* (Vol. 2008, pp. 699-717).

Schlenker, D.E. & Coles McKinnon, N. (1994). Assessing faculty performance using student evaluation of instruction. *U.S. Department of Education, Research/Technical report. Office of Research and Improvement.* (ERIC Microfiche ED 371 667).

Simpson EH. 1951. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society. Series B (Methodological) 13:238–241.

Stark, P., & Freishtat, R. (2014). An evaluation of course evaluations. ScienceOpen. *Center for Teaching and Learning, University of California, Berkeley*. https://www.scienceopen.com/document.

Treischl, E., & Wolbring, T. (2017). The causal effect of survey mode on students' evaluations of teaching: Empirical evidence from three field experiments. *Research in higher education*, *58*(8), 904-921.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185* (4157), 1124-1131.

Valencia, E. (2020). Acquiescence, instructor's gender bias and validity of student evaluation of teaching. *Assessment & Evaluation in Higher Education*, *45*(4), 483-495.

van der Lans, R. M., van de Grift, W. J., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, *50*, 88-95.

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in medicine*, *17*(1), 101-110.

Wilhelm, A. G., Rouse, A. G., & Jones, F. (2018). Exploring differences in measurement and reporting of classroom observation inter-rater reliability. *Practical Assessment, Research, and Evaluation, 23*(1), 4.

Young, M. E. (2018). A place for statistics in behavior analysis. *Behavior Analysis: Research and Practice*, *18*(2), 193.

Citation:

Howard, J. N.. (2022). Gauging Teaching Performance: Observational Sampling Opportunity, Reliability, and the Manifestation of True-Response Data. *Practical Assessment, Research & Evaluation*, 27(1). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/1/

Corresponding Author

Jeffery N. Howard
Northern State University
Aberdeen, SD, 57401 USA

email: jeffrey.howard [at] northern.edu