Analysing Student Evaluations of Teaching: comparing means and proportions

B.D. McCullougha* and Darrell Radsonb

aDepartment of Decision Sciences, Drexel University, CoB, 3141 Chestnut St, Philadelphia, PA 19104-2875, USA; bSchool of Business and Economics, Michigan Technological University, Houghton, MI 49931, USA

(Received 21 September 2010; final version received 3 July 2011)

Student Evaluations of Teaching (SETs) play a central role in modern academia. They are used for tenure, promotion, teaching improvement and other important decisions. One would think that the data collected from a SET would be analysed correctly, but such is typically not the case, as can be seen in this study later. Therefore we propose a correct method for analysing SET data. The present paper compares the two methods on a large data-set of actual SETs. We show that the traditional method can misrepresent a teacher’s performance, and that the traditional method can be extremely sensitive to outliers; neither of these characteristics is desirable. In contrast, the proposed method appears to suffer from neither of these defects.

Keywords: assessment; data analysis; statistics

1. A better method for analysing Student Evaluations of Teaching (SETs)

Student Evaluations of Teaching (SETs) play an important role in University life, especially in regard to tenure and promotion decisions, annual raises and for the improvement of teaching. Given their importance, one would think that great care and concern would be given to the proper analysis and interpretation of SET data. However, this is not the case. The widely used method for analysing SETs is statistically invalid, and near meaningless for their explicit and implicit applications: making discriminations between a professor and a standard (promotion and tenure decisions), or between professors (pay raises and teaching awards), or between an individual professor’s performance over successive classes (teaching improvement). It may well be true that the traditional method of analysing SETs would be valid merely for indicating extreme cases of ‘good’ or ‘bad’ teaching, but it is not sufficiently refined to permit the types of distinctions that SETs are typically used to make.

In Section 2, we briefly review the necessary ideas from the field of measurement theory to demonstrate the underlying problems with the traditional method of analysing SET data. Although this material may be known by most readers, it is included to allow this paper to be accessible to those who need to use SETs but have no statistical training. In Section 3, we review a standard application of the traditional method for analysing SET data and discuss how it violates basic statistical assumptions so that its results are, in large part, meaningless. In Section 4,
we present a new approach, taking care to demonstrate that it is a statistically
valid one for making the requisite types of discriminations, and compare it to the
traditional method using a large data-set of actual SET data. In Section 5, we present
a univariate analysis of the new approach using data from 49 classes taught to 737
students by 22 instructors. Section 6 contains a bivariate analysis using the same data
from Section 5. In Section 7, we outline a step-by-step approach to applying the new
method and a research agenda for investigating the new method.

Let us be clear: we are not suggesting that all schools immediately abandon the
traditional method since this may not be practical given years of tradition. We are
suggesting that our method is worth investigating and that some schools might wish
to implement both methods so that they may be compared in practice. In addition,
although there exist several sophisticated, statistically valid methods for analysing
SETs (Andrich, 1978; Göb, McCollin, & Ramalhoto, 2007; Tastle, Russell, &
Wierman, 2008), these methods are useful for academic research but not practical for
administrative purposes because without a graduate level knowledge of statistics,
these methods cannot be easily applied. We present a much-needed new method that
is relatively simple to apply and simple to comprehend, that satisfies the twin goals of
applicability and comprehensibility, and that is statistically valid.

2. Measurement theory

Definition (Sarle, 1997): measurement of some attribute of a set of things is the
process of assigning numbers or other symbols to the things in such a way that
relationships of the numbers or symbols reflect relationships of the attribute being
measured. A particular way of assigning numbers or symbols to measure something
is called a scale of measurement.

The theory of measurement ensures that our analyses of data are correct in a
meaningful way. The classic example of a meaningless statement is the weatherman’s
assertion that 60°F is twice as warm as 30°F; the ‘twice as’ applies only to the numbers
being used to measure the temperature, not to the temperature itself. We can say the
difference between 32°F and 96°F is 64°F, but we cannot say (meaningfully) that the
latter temperature is three times the former, even though it is true that 3 times 32
equals 96. To see this, consider the equivalent representation of these temperatures in
the Celsius scale:

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Measured in Fahrenheit</th>
<th>Measured in Celsius</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature A</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>Temperature B</td>
<td>96</td>
<td>35.5</td>
</tr>
</tbody>
</table>

We know that what is true in Fahrenheit must be true in Celsius. If the difference
between temperatures A and B is 64 F, we know that 64 F equals 35.5 C. Indeed, this
is true, just as 32 F = 0 C (freezing point) and 212 F = 100 C (boiling point).
However, it is not true that Temperature A is three times Temperature B, because
while 32 times 3 equals 96, it is not true that 3 times zero equals 35.5. It is important
not to confuse the properties of the numbers with the properties of the things being
measured. Just because 3 times 32 equals 96, it does not follow that a temperature of
96° is three times a temperature of 32°. Yet the prevailing method of analysing SET data does just this: it conflates the properties of the numbers of the Likert scale with the student’s response.

It is worth quoting from Sarle (1997):

When we measure something, the resulting numbers are usually, to some degree, arbitrary. We choose to use a 1 to 5 rating scale instead of a −2 to 2 scale. We choose to use Fahrenheit instead of Celsius. We choose to use miles per gallon instead of gallons per mile. The conclusions of a statistical analysis should not depend on these arbitrary decisions, because we could have made the decisions differently. We want the statistical analysis to say something about reality, not simply about our whims regarding meters or feet.

Measurement theory provides a way to ensure that the conclusion of one’s analysis doesn’t change if the unit of measurement is switched, say from inches to feet. The way to do this is to know what type of data you have and what types of statistics will preserve your answer when you change the unit of measurement. Stevens (1946) set forth the four types of data (nominal, ordinal, interval and ratio) and the methods of analyses that can be performed on each data type. These principles are part of the canon of the statistics profession and are routinely presented in elementary statistics texts.

Nominal data can be grouped into categories, e.g., the colours of cars: red, blue, green, yellow, etc. The categories cannot be ranked; it cannot be said that red is better than blue. A legitimate statistic that can be computed for nominal data is the mode, which is the most commonly observed value. Suppose the data consist of the religions of six persons: Episcopalian, Catholic, Buddhist, Baptist, Episcopalian and Mormon. The mode is ‘Episcopalian’. The average, or mean, cannot be calculated because the data cannot be summed: what is Catholic plus Buddhist? Assigning numbers to the religions does not enable calculation of the mean. If say, Catholic is recorded as ‘2’ and Buddhist as ‘1’ and Mormon as ‘3’, it is still true that 2 + 1 = 3 but it does not follow that Catholic plus Buddhist equals Mormon. Nor does it follow that the difference between Catholicism and Buddhism (2−1 = 1) is the same as the difference between Buddhism and Mormonism (3−2 = 1).

Ordinal data can be grouped into categories, and the categories can be ranked. Consider Olympic medals. We know that Gold is better than Silver, but not by how much. Similarly, Silver is better than Bronze, but we cannot say by how much. Again, we could report the type of medal by a number, e.g. Gold = 3, Silver = 2 and Bronze = 1, but we could not perform meaningful mathematical operations on these numbers and then apply the results of those operations to the categories of medals. Certainly we could say that 1 + 1 + 1 = 3, but it does not follow that three Bronze medals equals one Gold medal. Similarly, just because 3−2 = 1, it does not follow that the difference between a Gold medal and a Silver medal is exactly equal to a Bronze medal. A legitimate statistic that can be calculated for ordinal data is the median, which is the middle of an ordered set of data. For example, if the data are: 4, 2, 7, 6, 3, 8, 5, 1, 9, first place them in order (1, 2, 3, 4, 5, 6, 7, 8, 9) and then select the middle value (5).

Interval data can be represented on a numerical scale, and the difference between two points can be meaningfully calculated, but not the ratio of two measurements. The typical example is temperature. A very important consideration in the analysis of interval data is that, when comparing two measurements, only the difference between
them matters, not the actual values of the measurements. $A = 1$ and $B = 2$ is the same thing as $A = 11$ and $B = 12$ because in both cases, the difference is one. Another important consideration is that calculating the mean (average) is a permissible operation for interval data.

Lastly, ratio data can be represented on a numerical scale, and ratios can be formed: it is meaningful to say that one quantity is three times another. For example, length is a ratio scale measurement. If one item is three times as long as another in inches, this remains true when we change the scale to feet or metres.

Different types of mathematical operations can be legitimately performed on each type, as summarised in Table 1.

These principles are well known, taught routinely in introductory statistics classes and applied in data analysis in the quantitative disciplines, e.g. statistics departments, business and engineering schools. In what may only be attributed to contagious cognitive dissonance, these principles are forgotten when applied to the analysis of SET data. It is extremely common for the ordinal data in SETs to be treated as interval data, despite the fact that ‘[T]he effects of treating ordinal data as interval can be disastrous’ (Mayer, 1971, p. 519). Other problems in the traditional approach to analysing SETs are discussed at length in Langbein (2007).

A distinct minority of researchers contends that taking means of ordinal data is permissible, and usually cites a paper by Velleman and Wilkinson (1993). Two things are important to note when judging such claims. First, while Velleman and Wilkinson criticise Stevens’ categories, they never say that taking means of ordinal data is permissible. Second, these researchers’ arguments have convinced no one that taking means of ordinal data is valid. We can find no introductory statistics book stating that calculating means of ordinal data is a legitimate statistical operation that provides any meaningful interpretation of the data.

3. The traditional method of analysing SETs

Suppose that the ‘summary question’ (overall question) on an SET is, ‘The professor is a highly effective teacher’. We stipulate that all students have the same definition of ‘highly’ and all students have the same definition of ‘effective’. For example, it is not the case that one student thinks an effective professor ‘made me learn a lot even though I didn’t enjoy it’ while another student thinks that an effective professor is one who ‘made it fun to attend class’. If all the measuring devices, students in this case, are not measuring the same thing the same way, then the measurements cannot be aggregated. Simply assuming that they all measure the same thing the same way does not solve the problem.

Suppose that there are 30 students in the class and they all answer the survey as shown in Table 2.

Table 1. Statistics that can be computed on data types.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Permissible statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Mode</td>
</tr>
<tr>
<td>Ordinal</td>
<td>The above, plus median and percentile</td>
</tr>
<tr>
<td>Interval</td>
<td>The above, plus mean and standard deviation</td>
</tr>
<tr>
<td>Ratio</td>
<td>The above, plus coefficient of variation and logarithms</td>
</tr>
</tbody>
</table>
Clearly, the response categories (‘strongly agree’, ‘agree’, ‘neutral’, ‘disagree’ and ‘strongly disagree’) are ordinal. The traditional method of analysing SETs is to assign numbers to these ordinal categories, treating the data as interval data and computing the mean (average). Using the data in Table 2, the professor’s average score for the question is 115/30 = 3.83. It is important to ask, ‘3.83 what?’ What is the unit of measurement? There is none. What is being measured? If someone thinks that ‘agreement with the question’ is being measured, then the next time that person asks you, ‘Do you agree?’ you can answer ‘2.5’ and expect the listener to know precisely how much you agree. Of course this cannot be true because 3.83 is the putative mean of an ordinal variable.

Suppose the ‘average’ on this question for the entire department was 3.75. What can we infer about the professor’s performance of 3.83 in this course? In particular, can we infer that this did not happen by chance and that the professor really does do better, on average, than the typical member of the department? Frequently, the answer would be, ‘Yes, the professor exceeds the department average, because 3.83 is greater than 3.75’, but this is the untutored answer. Actually, to answer this question, a statistical test must be performed. Does this deviation of 0.08 on the high side provide strong evidence that the professor is a superior performer in the classroom and would continue to outperform his colleagues? A typical, but incorrect, way to proceed would be to treat the 3.75 as a population average and then perform a one-sample test of means. To perform this test, the standard deviation of the professor’s scores must be calculated, which for these data is 1.1. To be significantly different, the professor would need a mean score of greater than 4.16 or less than 3.34 (the interval is given by 3.75 ± 1.96*1.1/√30). Consequently, even if taking means of ordinal data were legitimate, there would be no evidence that the professor is better than the department average in this particular case.

Of course, none of this is valid because taking means and calculating standard deviations of ordinal data are meaningless. It is also common practice for university administrators to rank their faculty according to a summary SET question, drawing a couple of lines through it and thus divide the faculty into ‘good’, ‘average’ and ‘bad’ teachers. For such an exercise to be valid, statistical methods should be employed, and the statistic used should be legitimate for ordinal data; the method we advocate in Section 4 is one such method.

Due to the prevalent use of the traditional method, we deem it necessary to provide further evidence that SET data are not interval data. First, we observe that many schools use different response categories for SETs. Searching the Internet, we easily found several different categorisations, as shown in Table 3.

Anyone who believes that School A employs a legitimate interval data Likert scale should react with shock and abhorrence to the necessarily illegitimate scales

<table>
<thead>
<tr>
<th>Response</th>
<th>Number of students</th>
<th>Assigned number</th>
<th>Total points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly agree</td>
<td>9</td>
<td>5</td>
<td>45</td>
</tr>
<tr>
<td>Agree</td>
<td>12</td>
<td>4</td>
<td>48</td>
</tr>
<tr>
<td>Neutral</td>
<td>5</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Disagree</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td></td>
<td>115</td>
</tr>
</tbody>
</table>
employed by the other schools. School A contends that the distance between
‘strongly agree’ and ‘strongly disagree’ equals three and, if this is an interval measure,
this difference is not a matter of opinion and the question is not open to debate: the
distance between ‘strongly agree’ and ‘strongly disagree’ cannot equal 4, 5 or 6 as
Schools B, C and D contend, respectively. The fact that nobody sees the big
difference between A, B, C and D means that everybody knows that these are ordinal
data, not interval data.

4. A better method for analysing SETs
Statisticians have built up an impressive and solidly grounded collection of methods
for analysing data that are expressed in the form of proportions. To make use of these
methods, all that is necessary is to re-express the SET data in the form of
proportions. There are many ways to do this, below we present a few.

Consider the example presented in Table 2 in the previous section. The following
valid statements can be made:

(1) The proportion of students who answered ‘strongly agree’ is: $9/30 = 0.30$.
(2) The proportion of students who answered ‘agree’ or ‘strongly agree’ is:
$21/30 = 0.70$.
(3) The proportion of students who answered ‘neutral’ or ‘agree’ or ‘strongly
agree’ = $26/30 = 0.87$.

Which proportion is relevant may depend on the school and its particular goals. If
the goal is to identify the ‘very best’ teachers for the purpose of giving a teaching
award, then (1) would be useful. If the task is to promote a professor on the basis of
‘good’ teaching, then (2) might be useful. However, if the faculty were small and
filled with outstanding teachers, then (2) might set the bar too high, and (3) (or some
entirely different proportion) might be appropriate.

Suppose that a goal, based on taking the mean of ordinal data, is set for an
individual faculty member, or for a department or even for a school. For example, we
know of one university which requires each department head to meet with all faculty
whose class mean for a particular question on their SET is below 3.0, since an average
score below 3.0 is considered indicative of ineffective teaching. The goal is arbitrary
and is devoid of interpretation beyond the numeric ‘3.0’. Contrast this with the
following goal based on proportions: ‘We want our faculty to have 70% of students
answer “agree” or “strongly agree” to the summary survey question’. Everybody can
easily understand what this means. Goals set in terms of proportions are more readily
comprehended than goals set in terms of the means of ordinal variables.
Some persons understand that a mean score of 3.0 has no individual interpretive meaning for a particular class or instructor, yet mistakenly believe that it can be validly used for relative comparisons between and among faculty. That is, the means of ordinal data can be used to rank faculty with respect to their teaching ability. As a result, many schools rank faculty by merely sorting the mean scores from a particular question on the SET from high to low. Some schools even set cut points from these rankings to find the teachers at the top and the teachers at the bottom who need assistance or guidance to improve their teaching. Ranking mean scores, or creating categories from these means, and believing that a higher mean score is actually better than a lower mean score is statistically invalid. To make such judgements, statistical ranking techniques for means would have to be employed to determine significant differences (if any) between the rankings. This is beyond the knowledge of most administrators and faculty. In any event, since the mean of ordinal data is not a valid metric, as we have shown, this behaviour of ranking by sorting the means should cease.

The use of proportions can bring the focus more on the individual class performance of an instructor rather than some arbitrary ranking of mean scores. Appropriate goals can also be determined for particular SET questions (In general, if a goal cannot be determined, the obvious questions are, why are we asking the students to respond to that particular SET question, and why do we bother to collect the data?). For example, a department may determine that an appropriate goal is for at least 80% of the students to respond ‘agree’ or ‘strongly agree’ to the SET statement: ‘The syllabus clearly outlined the grading requirements for this course’. This numerical goal should not be determined arbitrarily. Rather, it could be determined by: (1) an analysis of historical data to reveal the proportions reported in the past to identify if this goal is realistic, and (2) an administrative judgement that at least 80% is satisfactory or acceptable for our classes, departments or university (acknowledging that you are willing to accept at most 20% not responding ‘agree’ or ‘strongly agree’).

If a particular faculty member does not meet a particular goal, investigation should proceed without judgement into why this could have occurred. The reasons may be numerous: small class sizes where the response of one or two students can lead to excessive variation in the proportion; miscommunication between the faculty and students; the SET question is misinterpreted and subjective; or it is indeed an area where the faculty member is deficient and has an opportunity to make course changes to improve the proportions.

Using this type of analysis, individual faculty could set their own personal goals based on a subset of the SET questions. Using the above example, an individual faculty member may set a personal goal to have 90% of the students report either ‘agree’ or ‘strongly agree’ to a particular question or set of questions. Using SETs in this manner could change the use of SETs from being implemented solely in a top-down approach to a bottom-up approach where faculty are setting their own goals based on information from the SET they view as having an impact on student learning. The legitimate statistic (proportions) gives a different picture than the illegitimate and traditional statistic, the mean. Much more insight can also be gained from looking at the barplot of responses since for SET data, a single statistic does not provide sufficient information. Therefore, combined with the reporting of proportions, faculty and administrators need to see the barplot to be fully informed. In the twenty-first century, with the ease of computing and report generation, there is
no reason why administrators should not use both legitimate statistics (proportions) and barplots to assess faculty performance based on SETs.

5. Using proportions – univariate analysis

We obtained SET data from one particular school within a university for a particular term. After dropping all classes with fewer than 10 responses, we were left with 737 respondents enrolled in 49 sections taught by 22 instructors; the mean class size is 15 students. For illustrative purposes only, we will analyse the functional survey question: ‘Overall, how would you rate this instructor’s teaching?’ The possible responses are: [NA] Not Applicable [1] Poor [2] Fair [3] Good [4] Very Good [5] Excellent. Figure 1 shows the distribution of all student responses. The vast majority of responses are 4 (Very Good) or 5 (Excellent).

The first question to consider is this: what is the unit of observation? Is it the student or the class? Since teachers are evaluated on the basis of how well they teach to a class, not to individual students, we opt to use the class as the unit of observation. In order for statistics calculated for a class to be valid, the class cannot be too small. At a school where class sizes are very small, it may be preferable to use the student as the unit of observation.

The second question to consider is this: what shall be the statistic of interest? Should it be the proportion of students who rate the class a 3, 4 or 5 (p345), the proportion of students who rate the class a 4 or 5 (p45) or the proportion of students who rate the class a 5 (p5)? We cannot answer this question; it depends on the purpose and use of the SET. It may even be desirable to focus on p12 (the proportion of students who rate the class a 1 or 2) if an administrator wants to identify teachers in need of some form of remedial help.

When using a statistic for discriminatory purposes, it is desirable that the statistic have more variance rather than less – if everyone has the same score (i.e. no variance), then there is no basis for segmenting the data. With this thought in mind, we look at the histogram of p345 given in the top left panel of Figure 2. This distribution is very skewed: almost all classes have 95% or more responses of 3, 4 or 5, so this statistic will not be very useful in discriminating between classes (and the instructors thereof). SET data are usually skewed. Of course, it may not be the case

![Histogram of responses to summary question.](image)
that an administrator would want to discriminate between the bulk of the teachers. If almost all the teachers are rated at least ‘good’, as this figure shows, then the administrator might well wish to focus on the few classes that are not doing so well – we will show how to do this later.

Examining the top right panel in Figure 2, we see that the bulk of the classes have 80% or more of the students responding ‘very good’ or ‘excellent’. Again, this distribution is very skewed. It shows, as the previous figure did, that most of the students appear to be quite happy with most of their instructors.

The bottom panel in Figure 2 shows the distribution in the proportion of students who rate the class ‘excellent’. This distribution is not skewed, and appears almost bell-shaped. This statistic, p5, would be an excellent choice for discriminating among instructors to possibly select the ‘best’ teachers.

6. Using proportions – bivariate analysis

Let us now compare the use of the proportional statistics p345, p45 and p5 to the ‘mean’ score of the ordinal data. We do so by using a traditional scatterplot. For illustrative purposes only, we put 95% confidence intervals about the average responses; we are not advocating the use of confidence intervals for making discriminations between or among teachers. In fact, we strongly condemn the practice, because the width of the interval depends on the number of classes. For example, as the number of classes increases, the width of the interval decreases. Hence, in a large department with many courses taught, the interval would be small and most teachers would be either above or below the interval. Nonetheless, for illustrative purposes, let us say that a class above the upper limit is ‘good’, a class within the limits is ‘average’ and a class below the lower limit is ‘bad’.

Since the mean of the ordinal data is unitless, there is no basis for an administrator to say that 3.5 is ‘good’ and 3.4 is ‘bad’. Confidence intervals are a way to establish bounds. Even then, no one knows what the bounds mean for this application because no one knows the answer to the question, ‘3.5 what?’ For proportions, this is not true. We know how to interpret the number 70% when we are told, ‘70% of the students responded “good” or “very good”’.

From our Figure 2 we know that the p345 statistic will not provide much discrimination and, indeed, it does not. Figure 3 shows a scatterplot of p345 and the SET question mean for each of the classes, where each data point is represented by the class number. Looking at Figure 3, we see that the bulk of the data are at the very top where the proportion equals 1.0, and the relation between p345 and the mean is decidedly non-linear. There is little agreement between the two statistics. The teachers considered ‘average’ by the mean (that is, falling between the lower and upper limits about 4.29) are considered ‘above average’ by the proportion. It is not the case that the mean and the proportion are just two ways of looking at the same thing. It is true that they will agree at the extremes. For example, if all the students rate a teacher very highly, both the proportion and the mean will be high. Similarly, if all the students rate a teacher very poorly, the proportion and the mean both will be low. When the students do not all register similarly extreme responses, the mean and the proportion will diverge, and this is where most cases fall. As noted previously, in such a case we know how to interpret the proportion but not a mean of ordinal data. Below we provide an ‘obvious’ example of how the usual interpretation of the mean of ordinal data is incorrect.
Figure 2. Histograms of statistics p345, p45 and p5.
For now, some noteworthy observations can be made by looking at where the mean and p345 differ in their assessments of specific classes. Consider Class 55, near the top left. Its mean score is 3.86 which is appreciably less than the mean of all the classes, 4.29. It is far outside the 95% confidence interval [4.15, 4.44]. Any administrator looking at this would conclude that the teacher of Class 55 did a much worse job than his colleagues. Such a conclusion would be incorrect.

According to the proportion and its confidence interval, the teacher of Class 55 is at least as good as most of his colleagues. What is the truth? To get to the truth, we have to look at the histogram for this course, which is given in the left panel of Figure 4. There are no dissatisfied students, that is no student responses of 1 (poor) or 2 (fair).

Compare Class 55 with Class 15, whose mean is 4.00. Class 15 has a higher mean score, but does have dissatisfied students, whereas Class 55 has no dissatisfied students. Is the mean a reasonable indicator of teacher performance? The class sizes are small, so even a few bad student evaluations have an appreciable impact, but the point is clear: the use of means and proportions can lead to different conclusions, and the former is more likely to be misleading since it is statistically invalid as we discussed earlier.

It is important to recognise that, given the emphasis placed on the use of means to interpret SET data, almost anyone looking at Class 55 would observe that it was more than two standard deviations below the mean for all classes and conclude that the teacher was doing a poor job compared to his peers. As the histogram reveals, nothing could be farther from the truth.

We turn next to the p45 statistic, which is plotted against the mean in Figure 5. The relation is much more linear, indicating more agreement between the two measures. Still, there is a number of discrepancies and some of them are worth exploring in detail. The confidence interval lines divide the figure into nine boxes. See the middle box on the right. These courses are rated as good according to the mean, but average according to proportion. Similarly, the middle box on the left is rated

---

Figure 3. p345 vs. $\bar{X}$, with 95% confidence intervals.
Figure 4. Histograms for Classes 55 and 15.
bad according to the mean, but average according to the proportion. The top box in
the middle is rated average according to the mean, but is good according to the
proportion.

Consider Classes 15, 16 and 40, all of which are rated as ‘average’ according to
p45, but which have different ratings according to the mean: ‘bad’, ‘average’ and
‘good’, respectively.

Is it the case that Class 16 is really that much better than Class 15? Examining the
individual class histograms in Figure 6, we see that Class 15 has a single response
equal to ‘1’ and a single response equal to ‘2’, but more ‘5’ responses than Class 16.
The low and high student responses should be considered collectively when judging
teacher performance but they are distorted by the computation of the mean. The
mean for Class 16 is almost exactly equal to the overall mean, while Class 15 is
3.9 standard deviations below the overall mean. Any administrator looking at
the means, especially one who is considered ‘statistically savvy’ enough to use a
confidence interval, would erroneously conclude that the instructor of Class 15 is
grossly deficient.

We have remarked previously that a problem with using means is that they are
susceptible to outliers, and the present comparison provides an excellent example of
this phenomenon. If the lone ‘1’ and lone ‘2’ responses are dropped from Class 15,
the mean for Class 15 rises to 4.31, above the overall mean. This is an example of the
mean being heavily influenced by the extreme responses. Simply by eliminating two
bad responses, the teacher of Class 15 could have gone from 3.9 standard deviations
below the mean to above the overall mean.

What about Class 40? Is it really that much better than Class 16? The histograms
show that the difference is stark: take most of the ‘4’ ratings from Class 16 and
change them to ‘5’ ratings; then Class 16 would be as good as Class 40. Yet p45
shows them both to be within the ‘average’ band. The reason is that the statistic p45
does not discriminate between exceptional and average teachers, at least not for this

Figure 5. p45 vs. $\bar{X}$, with 95% confidence intervals.
Figure 6. p45 for Classes 15, 16 and 40.
data-set. To explore this further, consider the scatterplot of p5 (the proportion of ‘excellent’ or ‘5’ responses) against the mean which is presented in Figure 7.

As can be seen, Class 40 is clearly above the upper limit for the proportion while Class 16 is clearly below the lower limit. The statistic p45 cannot distinguish between average and good student responses but the statistic p5 does because of its greater variability.

We need graphs like this for many more classes and schools to gain more experience and fully understand how best to use them. They are most interesting. If you want to find the truly excellent teachers, this might be the graph you want. For example, in Figure 7, observe how many are classified as exceptional (24) according to the mean but how many fewer (14) according to the proportion.

Now let us turn our attention to identifying student responses at the negative end of the scale. To do so, we use the statistic p12, the proportion of students who respond with 1 (poor) or 2 (fair). The histogram for this statistic is given in Figure 8.

How should one identify classes where students report an unacceptable proportion of negative responses? This unacceptable proportion could be 10% responses of 1 or 2 or perhaps 20%? That is an administrative decision that can only be made by: (1) an analysis of historical data for similar classes to understand the proportion of responses that are typical, and (2) a decision that negative responses greater than a particular number are unacceptable for the particular university or department. For the present data-set, the average class size is 15, and if even two students give a rating of either 1 or 2, then 2/15 = 13.3%. Considering this and the analysis of historical data, we might conclude that p12 > 0.15 is unacceptable for a particular class. Note, however, that for such small sample sizes (classes that are typically 15 or less), it could happen that three students give ‘fair’ or ‘poor’ responses for a particular teacher. In such a circumstance, having three students give fair or poor ratings repeatedly might be a sign of trouble, but if it only happened once it could be interpreted as random occurrence.
Examine the scatterplot of p12 against the mean, given in Figure 9. The use of the mean and its confidence interval would incorrectly label Classes 15 (which we examined earlier), 14, 11, 5 and 55 (which we also examined earlier) as problem teachers, since anyone falling outside the confidence interval is sure to attract attention. But look at Class 55, it has zero poor or fair ratings. Could this really be ‘bad’ teaching? It certainly could be judged so, according to the mean, but not according to the proportion$^{3}$

Figure 9. p12 vs. $\bar{X}$, with 95% confidence intervals.
7. Conclusions, applications and directions for further research

Since the use of proportions is statistically valid, and the use of means is at best questionable, from an administrative, academic and legal point of view, it is best to use proportions. Although the consistent use of proportions and eliminating the use of means may require (1) brave university administrators to take on the task of SET analysis reform, and (2) a cultural change on the interpretation and use of SETs by both administrators and faculty, we understand that these changes will take considerable time and energy but the long-lasting benefits are worth the effort. Therefore, we present recommendations as a start down the road to change:

(1) Review your current SET and ensure the scale used is consistent with solid social science research methods employing Likert scales. The few universities already down this path have consulted with their faculty experts in this area and use scales similar to the School B example in Table 3. This first step is important because if the definition of the measurement scale is judged inferior by many, any analysis of the data will be suspect. If however, this is determined to be a monumental exercise that may take years on your campus, eliminate this step for now and proceed for the meantime to the following step.

(2) Review your current SET and select a subset of questions to focus on. We acknowledge that SET question development could be the sole subject of another research paper. Briefly, we believe that well developed questions (1) are interpreted similarly and unambiguously by students, and (2) seek input from students in areas where they have the ability and knowledge to provide a judgement. We have already discussed point (1) in Section 3. Regarding point (2), many argue that students do not have the ability to judge if someone is an ‘effective teacher’ but few argue that students cannot judge whether the ‘syllabus clearly outlined the grading requirements for this course’. Therefore, select a subset of questions that can be objectively evaluated by knowledgeable students rather than subjective questions with multiple interpretations. (The ‘overall’ or ‘summary’ questions on most SETs are the ones that are highly subjective and, unfortunately, used most by administrators but most disliked by faculty. We recommend this question should ultimately be avoided.)

(3) For each question in the subset, ensure the data analysis (for each class or set of classes) provides the category response frequencies, the category proportions and the barplots as we have shown in this paper. SET question means may also be provided only as a point of comparison with the proportions as we wean ourselves off the means.

(4) When reviewing SET question data, and if the mean score is reported, ask yourself what information the mean score is providing and review how the analysis of the means has been used to make decisions. We believe that if one fully attempts to answer questions such as, ‘Is a mean score of 3.8 (on a five-point scale) satisfactory for our faculty?’ or ‘Is an instructor whose mean score is 3.8 doing a good job?’, the limited value and misinterpretation of means will become more apparent. This combined with a reread of Section 2 of this paper may go a long way in the realisation that means should not be used. It may be a good time to review how your school has used the means to
judge faculty teaching ability. For example, a school may realise that in the
past a mean score of three or better for a particular SET question (on a four-
point Likert scale) was determined to be satisfactory, good or even very good.
However, for particular classes, close observation of the proportions may
indicate that 15–20% of the students either responded ‘disagree’ or ‘strongly
disagree’ on the questions which had mean scores greater than 3.0. Is 15–20% of
the student body reporting negatively really satisfactory, good or even very
good? That is for each administrator or department to judge but if we just use
the means, this information on reported disagreement is lost.

(5) Determine a proportion (percentage) that is an appropriate goal for a class or
instructor within your university/department. For example, on a five-point
Likert scale, you may be interested in the proportion of students who
responded ‘strongly agree’ or ‘agree’ to a particular question. Alternatively,
you may be interested in the proportion of students who responded ‘strongly
disagree’ or ‘disagree’ to a particular question. The numerical goal should not
be determined arbitrarily but based on historical data for similar classes and
administrative judgement. Some faculty may wish to set their own personal
goals for particular classes and use the accomplishment of these goals in
annual faculty reviews.

(6) Track the proportions to the subset of questions over time. This can be done
using historical data (if available) or by starting with the current period and
going forward. This tracking could be done for a particular faculty member,
the same class (taught by multiple faculty) or for a grouping of similar classes.
Observing the data over time can determine if the faculty performance or the
department is improving over time. This can be done without statistical tests
only if the analyst does not overreact to every observed difference from
proportion to proportion over time since variation in these measures is
expected. In more quantitatively oriented departments, statistical tests could
be undertaken, as we have shown in the previous section, to isolate cases
where the analysis of the proportions provided different conclusions than the
use of the means (of course, there are no valid intervals for the means, but
there are for the proportions). The recommendations above could be
undertaken first as a pilot project in a department, preferably a quantitatively
oriented department, whose members do not have to be persuaded that the
traditional method of averaging ordinal data is invalid. The lessons learned
from their implementation should then be communicated to other depart-
ments. Questions that should be addressed include:

- Are teachers rated or judged as good (or bad) before using the means still
  viewed as good (or bad) using the proportions?
- What about the teachers who were in the middle before (i.e. not good or bad)?
  Do the proportions isolate ratings that are now viewed as good (or bad) since
  the proportions are able to provide easily interpreted data and discriminate
  between teachers better than an overall mean?
- How have individual faculty used the proportions to assess their teaching?

The data from such a pilot could be used to educate administrators and other
faculty. Such dissemination of results could lead to eventual rollout on a university-
wide basis.
Applications of the use of proportions could form the basis of research projects undertaken by an individual campus and throughout academia. Case studies of this nature would provide invaluable information on how SETs can be used to constructively evaluate teaching. Given the pervasive use of SETs, it is important to begin this analysis to foster discussion and change.

Although this paper is focused on the analysis of SET data, further discussion and research is needed on the types of questions or statements that should be included in an SET. Rifkin (1995) noted that ‘There is one criteria [sic] that most researchers agree on: student evaluations should be non-normative’, i.e. questions should compare teachers to an objective standard, not to each other. We stated previously that SET questions should be interpreted similarly and unambiguously by all students and this is a research question on the reliability of particular questions or statements in an SET.

But a larger question needs to be addressed: ‘Are students qualified to rate their instructors and the instruction?’ The answer is: ‘It depends on what questions are asked and how they are phrased’. For example, many would argue that ‘The teacher is knowledgeable in the subject matter’ is a poor statement for students to agree or disagree with, since students do not have the requisite subject-matter expertise to make this determination. Objective statements about the course syllabus or other course materials within the teacher’s control may be constructed and tested to ensure they are interpreted similarly and unambiguously by all students.

A much deeper controversy and research question is whether SETs, or individual questions within SETs, can measure ‘teaching effectiveness’. The current use of SET data by many administrators implies that ‘teaching effectiveness’ is being measured but many researchers would not agree and this discussion and research needs to continue (see, for example, Marsh & Roche, 1997; Sproule, 2000). Our larger point is this: there is no added value in correctly analysing SET results if the SETs themselves are asking the wrong questions.

Acknowledgements

We thank the referees for numerous comments and suggestions, which greatly improved the paper. We thank also Sheila Vaidya and Rodger Collons for useful conversations, and Tom McWilliams for comments. Finally, we are deeply indebted to David Urias, who procured the data-set we used, and who urged us to complete this paper; without his support this paper would not have been written.

Notes

1. We do not advocate the use of such an overall student assessment question to evaluate anyone’s ability to teach. This type of summary question is highly subjective, not interpreted uniformly by students and, unfortunately, used extensively while most disliked by faculty. We recommend this question be ultimately avoided in making administrative and merit (pay raise) decisions and the focus turn to more objective questions that usually exist on a well-designed SET.

2. In fact, for the statistically inclined, 3.86 is 5.7 standard deviations below the mean.

3. How to implement goals in practice is not obvious, and here we raise just one issue that needs to be worked out. Suppose that a school’s goal for satisfactory performance is that 70% of respondents judge the teacher to be ‘good’ or ‘excellent’ (4 or 5 on a five point scale). Does the teacher meet the goal if a test of $H_0: p \geq 0.70$ vs. $H_1: P < 0.70$ fails
to reject the null, or should the test employ $H_0: p_{45} \leq 0.70$ vs. $H_1: p_{45} > 0.70$? Should the null and alternative hypotheses be structured the same way for ensuring satisfactory performance and for identifying outstanding teachers?

References


