
This is the postprint version.

This is an Accepted Manuscript of an article published by Taylor & Francis in December 2012 in *Assessment & evaluation in higher education*, available at: [http://www.tandfonline.com/10.1080/02602938.2011.592935](http://www.tandfonline.com/10.1080/02602938.2011.592935)

©2012, Taylor & Francis

Reproduced by Deakin University with the kind permission of the copyright owner.

Available from Deakin Research Online:

[http://hdl.handle.net/10536/DRO/DU:30049593](http://hdl.handle.net/10536/DRO/DU:30049593)
The performance of a student evaluation of teaching system

Stuart Palmer

Institute of Teaching and Learning, Deakin University, Geelong 3217, Australia

Student evaluation of teaching (SET) is now common-place in many universities internationally. While much effort has been devoted to examining the statistical validity of SET instruments, there has been limited examination of the methodological and consequential validity (together referred to as ‘utility’) of the ways in which SET data are used. This paper examines the SET system at Deakin University from the perspective of utility. It draws on publicly available SET results for an entire annual cycle of unit offerings. Consideration is given to the representativeness of the data produced, and to the utility of the data reported, by the system. While this investigation focuses on the SET system currently employed at Deakin University, it offers both an analysis methodology and conclusions that can be applied more generally.

Introduction

Student evaluation of teaching (SET) has a long history, has grown in prevalence and importance over a period of decades, and is now common-place in many universities internationally (Davies et al. 2007; Hulpiau et al. 2007; Kember et al. 2002; Lemos et al. In print; Smith 2008). SET data are collected for a range of purposes, including: as diagnostic feedback to improve the quality of teaching and learning; as an input to staff performance management processes and personnel decisions and in response to government statutory requirements (Kember et al. 2002; Neumann 2000). A recent national investigation into SET practices in Australian higher education found them to vary widely, to be largely idiosyncratic, to have been developed internally without reference to each other, and to rarely
have an explicit theoretical basis (Barrie et al. 2008). Not only was there wide variation found in the items making up SET instruments, but also in the SET administration and reporting processes. It is important to distinguish between the SET instruments themselves, for which various claims of validity and reliability might be made or demonstrated, and the related processes of administering the instruments, the reporting of the SET data thus obtained and uses to which the data are applied. Elsewhere it has been noted that while much effort has been devoted to examining the statistical validity of SET instruments, there has been limited examination of the methodological and consequential validity (together referred to as ‘utility’) of the ways in which SET data are used (Beran et al. 2005). Additionally, it is noted that even well founded and long-standing SET instruments and/or processes may need to be re-examined if external environmental factors and/or internal institutional goals and objectives change over time (Oliver et al. 2008).

Rather than presenting a case study of the psychometric properties of another SET instrument, this paper examines the Student Evaluation of Teaching and Units (SETU) system at Deakin University from the lesser investigated perspective of utility. It draws on publicly available SETU results for an entire annual cycle of unit offerings that represent 74498 sets of SETU ratings and 188391 individual student unit enrolments. Consideration is given to the representativeness of the data produced by the SETU system and to the utility of the data reported by the SETU system. While this investigation focuses on the SET system currently employed at Deakin University, it offers both an analysis methodology and conclusions that can be applied more generally.

**Deakin University’s Student Evaluation of Teaching and Units survey**

In Australia, Deakin University is a major provider of distance and online education. In addition, it teaches on-campus at four campuses located in three cities in the State of Victoria,
with campuses spanning metropolitan, regional and rural locations. Deakin University currently teaches on a trimester system, with three teaching periods per year of equal duration and status. In total, approximately 39,600 students are enrolled in studies (Deakin University 2010b). Deakin University’s ‘Quality Assurance and Continuous Quality Improvement in Respect of Academic Matters’ Procedure (Deakin University 2003), requires that its ‘Evaluation of Teaching and Units’ Procedure (Deakin University 2009) be enacted. For all higher education award courses, except higher degrees by research, the latter procedure requires that:

- unless a case is made for exemption, the SETU questionnaire is administered to students enrolled in every unit of study every time it is offered;
- for staff involved in teaching, SETU results form part of their performance review process; and
- summary SETU results are made available for query on a publicly accessible reporting website (Deakin University 2010c).

The SETU instrument, as a standardised, centrally administered questionnaire, was first introduced in 2003, and its current form was introduced in 2006. It consists of ten core question items:

1. This unit was well taught.
2. The course materials in this unit were of high quality.
3. The workload in this unit was manageable.
4. Requirements for completing the assessment tasks in this unit were clear.
5. The teaching staff gave me helpful feedback.
6. The library resources met my needs for this unit.
7. I would recommend this unit to other students.
8. The technologies used to deliver the online content in this unit performed satisfactorily.
9  The on-line teaching and resources in this unit enhanced my learning experience.

10  This unit challenged me to learn.

SETU respondents rate the core question items on a five point scale (1=strongly disagree; 2=disagree; 3=neutral; 4=agree; 5=strongly agree) with a ‘not applicable’ option included.

The publicly accessible SETU reporting website (Deakin University 2010c) allows anyone with an interest to query the results for the ten core SETU items, based on a selection of evaluation period, faculty, school, unit and student enrolment location. The data reported for each unit includes total enrolment, number of responses and computed response rate for the enrolment location(s) selected. Additionally, for each unit, for each of the ten core SETU items, the number of responses, mean rating, standard deviation of the mean rating, percentage agreement, percentage disagreement and percentage difference are reported. The percentage agreement is based on the number of responses in the agree and strongly agree response points, percentage disagreement is based on the number of responses in the disagree and strongly disagree response points, and the percentage difference is computed from the percentage agreement less the percentage disagreement. SETU results are publicly reported for a unit unless the number of responses is less than ten; the presumption being that anything less than ten responses is an unrepresentative sample size. Individual teaching staff can login to the reporting website, and once authenticated, can access additional SETU information that applies specifically to them, i.e., student ratings and comments regarding their individual teaching roles.

On the face of it, this system for evaluating teaching and units, including supporting policies, implementation procedures and reporting of results, would appear to be comprehensive, transparent and informative. The following section presents an analysis of this system, based on publicly available information. The analysis covers a whole year period, including trimester 2 2009, trimester 3 2009/2010 and trimester 1 2010.
Analysis of the units included in the SETU system

The list of higher education coursework units of study offered by Deakin University is given in the online Course Handbook (Deakin University 2010a). For the period under consideration a total of 2452 separate unit offerings are listed. For the same period, the publicly accessible SETU reporting website (Deakin University 2010c) contains records for 1989 of these unit offerings. This means that 463 (18.9 percent) of the listed unit offerings were not included in the SETU process, presumably exempted on some grounds. Additionally, a further 557 (28.0 percent) of the units with a record in the SETU reporting website did not actually publicly report any results due to the number of respondents being less than ten. So, in total, for the period under consideration, only 1432 unit offerings, or 58.4 percent of those listed in the handbook, actually publicly reported any SETU data. The analysis that follows necessarily focuses on the data set of units that provided information via the public SETU reporting website – no data are publicly available for non-reporting units.

Figure 1 shows the distribution of numbers of respondents for those units that did report. The distribution is skewed to low numbers of respondents, with the mean number of respondents being 52.0 and the median number of respondents being 29. If it is considered that there are an additional 557 non-reporting units with number of respondents less than ten, then the overall respondent distribution for units included in the SETU system is highly skewed to low numbers of respondents.
Figure 1. Distribution of numbers of respondents for reporting units.

For those units that did report, a key question is how representative is the respondent sample? In this regard, response rate is an important indicator. Figure 2 shows the distribution of response rates for units reporting via the SETU. The distribution is approximately normal (Kurtosis = 0.51; Skewness = 0.57), with a mean response rate for reporting units of 43.0 percent.

Figure 2. Distribution of response rates for reporting units.
Another important indicator of sample representativeness is how closely the sample (students in a unit responding to SETU) matches the population (the entire student enrolment for a unit) on key demographic dimensions. In the case of student unit enrolment, important demographic characteristics might include age, gender, mode of enrolment, home campus, enrolled program of study (where students from more than one program enrol in a unit), etc. In the data provided by the SETU reporting website, the only demographic information available is the numbers of enrolled and responding students at each of the five campus locations (Burwood, Geelong, Warrnambool, Waterfront and off-campus). Of the 1432 units that reported SETU data, 606 (42.3 percent) have students enrolled at a single campus only; meaning that the principal measure of respondent sample representativeness resides with the response rate. For the balance of reporting units, it is possible to perform a contingency table analysis to test whether the proportions of respondents from each campus are similar to the overall unit enrolment proportions by campus location. In all cases, it was possible to perform an exact, two-sided version of Fisher's test to assess the representativeness of the respondent sample on the basis of enrolled campus. Based on a statistical significance level of 0.1 (i.e., the observed sample could be expected to occur at least one in ten times by chance) a majority of respondent samples would be judged as adequately representative on the basis of enrolled campus. However, there were 33 units for which the observed proportions of respondents, on the basis of enrolled campus, would not be considered an adequately representative sample.

The question of what is an acceptable response rate is not straightforward to answer, nor is there general agreement on the desirable minimum response rate. A high response rate is desirable in as much as it contributes to the representativeness of the response sample and hence, lowers the likelihood that non-respondents would have provided a response
significantly different from the respondent sample. A lower response is acceptable if there is good evidence that the sample remains representative of the enrolled population. With only demographic data relating to enrolled campus provided by the SETU reporting website, there is limited information on which to make a judgement regarding the representativeness of the respondent sample, except perhaps to reject the reported data for a respondent sample that is clearly demographically skewed. With limited demographic information available, it is essential to achieve a good response rate to provide some confidence that the respondent sample is representative of the enrolled student unit population.

A minimally acceptable general response rate of 60 percent for surveys is indicated in the literature for research published in reputable journals (Baruch 1999; Timothy Johnson and Owens 2003). If a concession is made to that limit of ten percent (i.e., minimum required response rate of 50 percent), based on achieving a modest (Fisher’s exact test; \( p > 0.1 \)) demographic match based on enrolled campus, then 1066 of the 1432 units that reported SETU data would be rejected on the basis of response rate, for lack of confidence that respondent sample was adequately representative of the total enrolled unit population. A similar low rate of useful sets of SET data (23 percent, based on achieving a response rate of two thirds) is reported in the literature (Davies et al. 2007). There is some overlap in units reporting SETU data with low response rates and units with demographically skewed samples, so Table 1 provides an itemised summary of the analysis to this point. Table 1 shows that of the original 2452 individual unit offerings listed in the Deakin University handbook for the year spanning Trimester 2 2009 to and including Trimester 1 2010, only 363 units (14.8 percent of the total number of units) are: included in the SETU survey process; actually report results in the SETU system; have a representative demographic sample (on the basis of enrolled campus); and have a reasonable response rate (50 percent or greater).
Table 1. Components of the total number of units listed in the handbook.

<table>
<thead>
<tr>
<th>Category of units</th>
<th>Number of units</th>
<th>Percentage of total</th>
<th>Balance of 'good' units</th>
<th>Percentage of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total units listed in handbook</td>
<td>2452</td>
<td>100.0 %</td>
<td>2452</td>
<td>100.0 %</td>
</tr>
<tr>
<td>Units exempted from SETU</td>
<td>463</td>
<td>18.9 %</td>
<td>1989</td>
<td>81.1 %</td>
</tr>
<tr>
<td>Units not reported in SETU</td>
<td>557</td>
<td>22.7 %</td>
<td>1432</td>
<td>58.4 %</td>
</tr>
<tr>
<td>Units with skewed demographics</td>
<td>(33)</td>
<td>(1.3 %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Units with &lt;50% response rate</td>
<td>(1066)</td>
<td>(43.5 %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined rejected units</td>
<td>1069</td>
<td>43.6 %</td>
<td><strong>363</strong></td>
<td><strong>14.8 %</strong></td>
</tr>
</tbody>
</table>

While only unit response rate (including division by home campus where applicable) demographic information is actually reported by the SETU system, a rich set of demographic information is known about each respondent student. More of this demographic information could be reported in summary form by the SETU system to provide a clearer indication of the representativeness of the respondent sample obtained. If it was possible to determine that the respondent sample was representative on several demographic dimensions, the minimally acceptable response rate criterion could be lowered. Of course, the provision of more demographic information also permits additional dimensions on which the respondent sample might be called into question by being deemed as unrepresentative.

**Analysis of data reported by the SETU system**

Setting aside for the moment the representativeness of the units included in the SETU reporting system, the value of the data that are reported for units is now considered. As noted previously, in addition to unit enrolment and response rate statistics, this data includes the
number of responses, mean rating, standard deviation of the mean rating, percentage agreement, percentage disagreement and percentage difference for each of the ten core SETU items. Except for the limiting case where the response rate is 100 percent of the student enrolment (which did not occur in any of the 1432 units included in the SETU system during the period under consideration), the ‘mean’ rating reported via SETU is actually only an estimate of the true population mean rating. Even though the estimate status of the mean ratings is hinted at by the provision of the corresponding standard deviation of the mean rating, this important distinction is probably lost on most users of the SETU reporting system. These mean ratings provided are never qualified as estimates, and never presented with confidence intervals, in any reporting of SETU data, except for perhaps where individual staff attempt to compute confidence intervals for their own uses of their unit SETU data.

In addition to the mean respondent ratings and corresponding standard deviations, a number of extra data points are also provided for each SETU item rating. These include the percentage agreement, the percentage disagreement and the percentage difference. These derived data points are presumably provided to offer additional information for the analysis of evaluation ratings provided by students, and because the process of collapsing response scale data into dichotomous ratings is commonplace in survey research (Brody and Dietz 1997). This process of collapsing response scales into two categories is premised on an assumption that Likert-items produce separable dimensions of content (agree/disagree) and intensity (strongly agree to strongly disagree). However, it has been found that this is not universally the case, and that separability of these two dimensions should not be automatically assumed in the absence of results that would justify this operation (Duncan and Stenbeck 1987). It has also been noted that collapsing of scale categories effectively discards information provided by respondents on the intensity of their rating, and that the relationship
between a survey item and the response given may not be preserved under such collapsing (Sobel 1998).

In the case of the SETU survey instrument, which uses a five-point response scale, additional respondent information may be lost through the process of discarding any mid-point/neutral ratings in the computation of the agreement and disagreement percentage scores. During the period under consideration in this study, mid-point ratings for SETU items ranged from zero to 87.5 percent of the total number of responses recorded for an item in a unit, and the mean percentage of responses for an item in a unit that were discarded in computed agreement/disagreement/difference scores was 17.8 percent. This means that the percentage agreement, percentage disagreement and percentage difference scores are generally derived from fewer responses (and hence an even lower response rate) than their corresponding item mean rating estimates. In addition, the collapsing of the response scales means that there is not a bijective relationship (unique, one-to-one correspondence) between mean SETU rating and the three ‘agreement’ measures; i.e., it is possible to achieve a relatively high mean rating for an item while simultaneously achieving a relatively low percentage agreement, and vice versa.

However, concerns regarding the validity of collapsing response scales and the discarding of data in the derivation of percentage agreement, disagreement and difference scores are rendered somewhat moot because, as is shown in Figure 3, all three ‘agreement’ measures exhibit a strong linear correlation with their corresponding SETU item mean estimate. Figure 3 provides scatter plots that show the relationship between mean SETU rating and the corresponding computed percentage agreement, percentage disagreement and percentage difference for the 13564 individual unit SETU mean item ratings available via the SETU reporting website during the period under consideration. Table 2 quantifies this relationship by presenting the Pearson linear correlation coefficient \((r)\) and Spearman's rank
correlation coefficient (\(\rho\)) for each of the pair-wise combinations of mean SETU rating, percentage agreement, percentage disagreement and percentage difference. Table 2 reveals not only that all three ‘agreement’ measures are strongly and significantly correlated with their corresponding SETU item mean, they are also strongly and significantly correlated with each other.

(b) 

(c) 

(a)
Figure 3. Scatter plots of mean SETU item rating versus a) percentage agreement; b) percentage disagreement; and c) percentage difference.

Table 2. Pair-wise correlation coefficients for mean SETU rating, percentage agreement, percentage disagreement and percentage difference.

<table>
<thead>
<tr>
<th></th>
<th>Mean SETU rating</th>
<th>Percentage agreement</th>
<th>Percentage disagreement</th>
<th>Percentage difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean SETU rating</td>
<td>r = +0.939</td>
<td>r = -0.859</td>
<td>r = +0.959</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ρ = +0.935</td>
<td>ρ = -0.805</td>
<td>ρ = +0.950</td>
<td></td>
</tr>
<tr>
<td>Percentage agreement</td>
<td></td>
<td>r = -0.779</td>
<td>r = +0.975</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ρ = -0.741</td>
<td>ρ = +0.987</td>
<td></td>
</tr>
<tr>
<td>Percentage disagreement</td>
<td></td>
<td></td>
<td>r = -0.912</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ρ = -0.867</td>
<td></td>
</tr>
<tr>
<td>Percentage difference</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: In all cases for the stated correlation coefficient, the number of samples is 13564 and the significance level is $p = 0.000$.

**Discussion**

Whether it is viewed from a teaching and learning quality assurance/quality improvement perspective, or from a teaching staff performance management perspective, a unit evaluation system achieves an effective inclusion rate of 58.4 percent of offered units cannot really be classified as comprehensive; at best, it offers a moderate coverage of units of study on offer. Many of the units not included in the system will be systematically missing due to being
purposefully exempted on some basis, or having a low enrolment (hence, unable or unlikely to meet the minimum number of respondents for inclusion in reporting). In this situation, the group of units that are included cannot even be described as a random selection for the purposes of producing an overall, institution-level measure of SET. If moderate minimum sample acceptance criteria are enforced for demographic representativeness and response rate, then the effective unit coverage rate offered by the SETU system further reduces significantly; to a mere 14.8 percent for the criteria suggested in this investigation. This result of such a low ‘actually useful unit coverage rate’ alone is a prompter for reconsidering the fundamental assumptions underpinning the design of the evaluation system.

Given that many reporting units have low numbers of respondents and/or low response rates, the presentation of unqualified estimates of mean ratings, if taken at face value in decision-making, could lead to erroneous conclusions – estimates of sample means should always be reported with appropriate confidence intervals. In the absence of easy access to the raw sample data, but with values for sample size and standard deviation provided, some staff may attempt to estimate confidence intervals for the mean ratings reported using the common $t$-distribution method (Rasmussen 1992). However, this approach is only valid when the data from which the mean is calculated come from a large population and are approximately normally distributed. Where the sample data departs from these assumptions, other methods exist for estimating confidence intervals for skewed and/or small samples and/or small class enrolments (Norman J. Johnson 1978; Keller 2009; Miller and Penfield 2005; Rasmussen 1992), but these methods typically require access to the raw sample data and involve more complicated computation. To ensure accuracy and consistency of results, appropriate confidence intervals on the mean ratings should be provided by the SETU system itself.
While the derivation and reporting of percentage agreement, percentage disagreement and percentage difference values might be common practice and appear to provide additional information for use in decision making, in practice, they may confuse rather than illuminate. As there is no bijective relationship between a mean rating and its corresponding ‘agreement’ scores, and as some criterion for ‘acceptable teaching performance’ are expressed in terms of mean rating and others in terms of percentage agreement, it is possible for a teaching staff member and/or a unit to simultaneously pass and fail the quality benchmark(s) for SET. Any confusion caused by the presentation of the ‘agreement’ scores is unnecessary, as these data are largely redundant, having a high and significant correlation with their associated mean score, and, in most cases, providing little additional information.

Even if the reporting of SETU were simplified to the presentation of mean scores with appropriate confidence intervals, this would still leave teaching staff, and other interested stakeholders, with the task of interpreting the results. An external review of teaching and learning at Deakin University in 2007 concluded, ‘The SETU instrument … should be accompanied by a clear statement as to how it can be used, and its limitations.’ (PhillipsKPA Pty Ltd 2007, 68) One long-standing model of providing guidance on the interpretation of SET data is the Rating Interpretation Guides (RIGs) system (Lemos et al. In print; Neumann 2000; Santhanam et al. 2000; Smith 2008). Although the specifics of various RIGs-style systems vary, the essential element is the provision of a norm-based set of benchmarks for the ranking or comparison of SET results. These benchmarks are based on a set of units of study that are similar in certain relevant respects – typically class size, class year level and discipline grouping – to the target unit.

**Conclusion**
Based on the analysis presented here, there are a number of opportunities to improve the SETU system at Deakin University. While the SET policy headline requirement is ‘every unit every offer’, an effective raw reporting rate of 58.4 percent of all units of study offered does not equate to a comprehensive teaching quality assurance/improvement system. For a SET system that is applied so relentlessly, much of the effort is effectively wasted. A case can be made for both expanding the coverage of the SETU system (i.e., not exempting so many units) and reducing the frequency of application – to reduce survey fatigue, potentially increase response rates, make room for other important institutional surveys and allow measured consideration and response to student evaluation feedback. A statistical/risk-based approach to periodic unit evaluation would probably result in a more systematic, valid and reliable system that leads to more effective changes for quality improvement in the longer term.

Of the units that do report via the SETU system, many are based on low response rates and/or demographically skewed samples, calling into question the value of the data thus reported. Additional respondent demographic data should be reported, so a more complete assessment of sample representativeness can be made, and perhaps the minimum acceptable response rates can then be relaxed. All data that are estimates of class population ratings should be presented with appropriate confidence intervals. The reporting of derived data (i.e., data inferred from other student ratings) that are largely redundant and potentially confusing should be avoided. The provision of SET results should be accompanied by guidance on their interpretation – most academic staff are not experts in statistics.

While much of the SET literature focuses on the psychometric properties of individual instruments, this paper investigates the issues relating to survey administration and results interpretation. Regardless of the psychometric integrity of an instrument, these more mundane factors may strongly influence the value of the data produced. Where SET data are
used to reward and/or punish academic staff, there is a need to ensure that the system is robust, valid and equitable. The analysis method employed here provides a model that others could use to critically evaluate their own SET systems.

References


http://theguide.deakin.edu.au/TheDeakinGuide.nsf/7264c32fe71924374a2566f3000a65de/2b619ef68fd4fc03ca256e830022ce0e.


http://theguide.deakin.edu.au/TheDeakinGuide.nsf/7264c32fe71924374a2566f3000a65de/c9cab3702f4e79e8ca2574180017b13a.
Deakin University. 2010a. Course handbooks. Deakin University.


Deakin University. 2010c. Unit evaluation system. Deakin University.


Table 1. Components of the total number of units listed in the handbook.

Table 2. Pair-wise correlation coefficients for mean SETU rating, percentage agreement, percentage disagreement and percentage difference.

Figure 1. Distribution of numbers of respondents for reporting units.

Figure 2. Distribution of response rates for reporting units.

Figure 3. Scatter plots of mean SETU item rating versus a) percentage agreement; b) percentage disagreement; and c) percentage difference.