Beyond Average: Contemporary statistical techniques for analysing student evaluations of teaching

Kirsty Kitto[a,b] and Cameron Williams[a,c] and Lyn Alderman[a,d]

[a]*Queensland University of Technology (QUT), 2 George Street, Brisbane, 4001, Australia*

[b]*University of Technology Sydney, 15 Broadway, Ultimo NSW 2007 Australia*

[c]*Newcastle University, United Kingdom*

[d]*Department of Social Services, Australia*

CONTACT

Email: kirsty.kitto@uts.edu.au

Connected Intelligence Centre

University of Technology Sydney, 15 Broadway, Ultimo NSW 2007 Australia

Beyond Average: Contemporary statistical techniques for analysing student evaluations of teaching

Student Evaluations of Teaching (SETs) have been used to evaluate Higher Education teaching performance for decades. Reporting SET results often involves the extraction of an average for some set of course metrics, which facilitates the comparison of teaching teams across different organisational units. Here, we draw attention to ongoing problems with the naive application of this approach. Firstly, a specific average value may arise from data that demonstrates very different patterns of student satisfaction. Furthermore, the use of distance measures (e.g. an average) for ordinal data can be contested, and finally, issues of multiplicity increasingly plague approaches using hypothesis testing. It is time to advance the methodology of the field. We demonstrate how multinomial distributions and hierarchical Bayesian methods can be used to contextualise the SET scores of a course to different organisational units and student cohorts, and then show how this approach can be used to extract sensible information about how a distribution is changing in time. We present a report designed to facilitate sense-making for the more complex statistical methodology that we propose, and demonstrate how it can be used to ensure that this more complex methodology is still appropriately used in decision making.

Keywords: student evaluations of teaching (SET); distributions; contextualisation; hierarchical Bayesian model; sensemaking

Measuring student satisfaction with teaching

Student Evaluations of Teaching (SETs) have been used as an evaluation mechanism for decades. A wide variety of different formats have been used, from verified scales supported by educational theory, to *ad hoc* questions that are deemed important by an organisation and chosen with no attempt to demonstrate validity (Marsh, 2007). The way in which SETs are used in an organisation also varies substantially. They can be used: as diagnostic formative feedback to improve teaching and learning; for personnel decisions based around teaching effectiveness; by students to select courses; for quality assurance purposes and public accountability; and to feed the ongoing research in the area (Johnson, 2000; Marsh, 2007). This range of possibilities means that any given institution might use its particular SET regimen in a number of different ways, some of which can have a significant impact upon the professional lives of academic staff.

While SETs were traditionally administered in a face-to-face format at the end of a teaching period, they are increasingly moving online into both formal and informal modes (Alderman and Melanie, 2012; Otto, Sanford Jr, and Ross 2008). This exacerbates a number of existing concerns about response rates (Zumrawi, Bates, and Schroeder, 2014), various forms of bias (Marsh, 2007; Spooren, Brockx, and Mortelmans, 2013), demographic effects (Macfadyen et al. 2015), and potentially negative correlations with desired learning outcomes (Braga, Paccagnella, Pellizzari, 2014). Overall, the field is contested, and practices are often further confused by the way in which this vast and often contradictory literature may be distilled to form, modify, or confirm an academic's existing biases about the validity or invalidity of SETs in their own institutional setting. A particularly convincing examination of the many myths surrounding SETs is provided by Aleamoni (1999), and references therein.

The above concerns remain largely academic as long as SETs are merely used to provide diagnostic feedback. However, in an era increasingly focussed upon performative measures of teaching quality it is essential that decision makers and evaluators make use of best practice methods to analyse the data that SETs generate. However, we frequently see outmoded or inappropriate strategies brought to this task. For example, a very common usage of SET data involves the extraction of an average score obtained for a particular class, for either an item or a collection of items (Abrami, 2001; Marsh, 2007). This often leads to an implicit comparison of courses across different organisational units. Such an approach is easy to apply, and so can be used to rapidly generate hypotheses about relative levels of student satisfaction. However, it also hides a vast array of contextual data that may be affecting these average scores (see e.g. Rienties and Toetenel (2016) for a large scale exploration of the way in which learning design can impact upon SET scores).

Even more problematic, it appears that few studies have been conducted to establish whether academic staff and university decision makers interpret average SET scores in an appropriate manner. A notable exception is provided by Boysen et al. (2014), who demonstrated that both academic staff and administrators use general heuristic methods to evaluate SET scores, rather than appropriate statistical principles (Kynn, 2008; O'Hagan et al., 2006; Tversky and Kahneman, 1974). Notably, Boysen et al. (2014) demonstrated that differences in averages small enough to be within a given margin of error significantly impact upon the assignment of rewards to teaching staff. This is no particular surprise. Even in fields dominated by the mathematical sciences there are many results showing that people do not interpret concepts like error bars and confidence intervals correctly (Krzywinski and Altman, 2013). There is no reason to expect that those interpreting SET results will be any better at performing what is known to be a difficult task.

In what follows, we will discuss the problems associated with common current practice in more detail using a typical institutional dataset. This will leave us in a position where we can propose more appropriate ways in which SET data can be analysed using contemporary statistical methods. We will illustrate the type of output that the new model generates, and consider how the resulting more complex reports might be simplified to facilitate rapid sense-making by staff who are less numerically literate. In summary, rather than providing yet another study that draws attention to issues of validity or bias in SETs, this paper will focus instead upon demonstrating that it is possible, and desirable, to use contemporary statistical methods when analysing them.

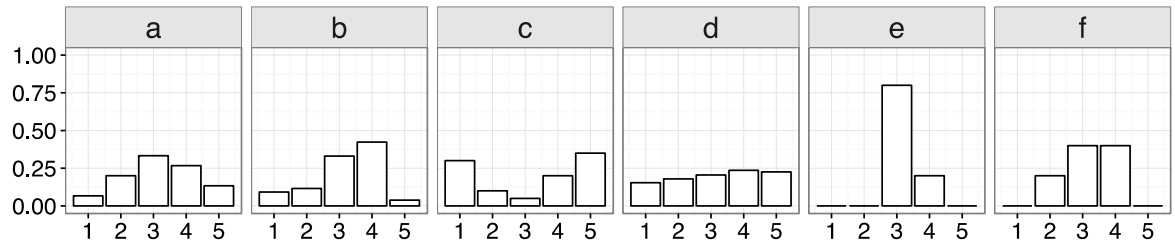

Figure 1. Six distributions obtained from Likert responses with a range of 1-5. Despite markedly different structures, each has an average of 3.2.

Pitfalls of current practice

Beyond the ongoing controversy surrounding the use of student satisfaction as a measure of teaching quality *per se*, there are a wide variety of *mathematical* reasons to be wary of the way SETs are often used in practice. Here we will focus much of our initial criticism upon concerns that arise from the use of average SET values in evaluation methodologies, before moving on to a discussion about wider problems concerning ordinal data, multiplicity, and the use of hypothesis testing.

Many distributions, one average

Many different distributions can lead to the same numerical value for an average SET score. Figure 1(a)-(f) demonstrates six ways in which different patterns of five-point Likert item responses can produce the same average score, in this case 3.2. Each distribution implies markedly different patterns of satisfaction in a class. The different patterns might be described as:

(a) Classically normal: This shape would be expected if SET responses arose from a homogeneous class of students, who cluster around a neutral response (e.g. the number 3 in Figure 1).
(b) Skewed: This response pattern is clearly skewed, indicating a shift away from the normal.
(c) Polarised: In this distribution we see a marked pattern where a significant portion of students are highly satisfied and another is highly dissatisfied.
(d) Flat: Each SET score is as likely as any other.
(e) Majority response: This pattern is common in smaller classes where the majority of students often select similar responses. In the illustrated case the majority vote is neutral, but this response pattern often occurs for classes with high scores.
(f) Non-polar cluster: As a cohort students are neither highly satisfied nor extremely dissatisfied with this course.

Note the dramatic difference in structure. Any academic who has been exposed to student evaluation data will quickly start to construct stories about what such distributions imply. For example, Figure 1(c) shows an extreme pattern that is more common than might be expected in university teaching; while a considerable portion of a class is highly satisfied, a second portion is extremely unhappy. How might such a wide polarisation arise? Often such SET signatures occur when the class contains cohorts from two distinct backgrounds. While one subset of students might be excelling, another may be lacking prerequisite knowledge and hence struggling. However, many other scenarios can lead to similar response signatures. Perhaps two different tutors have been engaged to teach into a large class, and one is obtaining far better satisfaction scores, a situation that would make it quite inappropriate to aggregate scores at the level of the whole cohort.

Figure 1 is only a selection of possibilities; there are a multitude of student response patterns which could still lead to the same numerical single-figure summary as an average. Are we to interpret them all the same way? This one-dimensional perspective would lead to the same action (e.g. intervention, if one were needed) for each class, despite the fact that each scenario is likely to benefit from different support. A metric reported as an average fails to draw our attention to this range of markedly different student satisfaction responses. Important information at the survey level regarding situational context has been lost, a problem which is further exacerbated when individual survey averages are aggregated to school, faculty or institution level (Rog, 2012). Using such an average value as a performance metric often results in a well justified outcry by academic staff.

Change in time

Changes in an average SET score over time have the potential to add another layer of obfuscation. What does an increase of 0.3 in the average for a SET item imply? Pedagogically this shift could arise for numerous reasons, but even from a measurement perspective, there are many ways in which a change in the distribution from one year to the next might result in the same shift in an average value. How are we to know what *form* of change in student satisfaction occurred? This becomes particularly important when we wish to disentangle the effects of factors such as transitions in teaching teams or changes in assessment from year to year.

A similar problem arises when we consider the way in which an average for a SET item might *not* be changing in time. As we saw in the above section, the same average score might hide a large amount of change in the underlying distribution of student responses. For example, a move in cohort satisfaction from the distribution depicted in Figure 1(d) to 1(e) could perhaps be regarded as an improvement (albeit at the cost of losing a few highly satisfied students), but this would not be discovered in an institution that was focussed upon reporting average values.

Decision makers are often trying to allocate limited resources to improve the student learning experience. Some courses might be underperforming when compared to the organisational context, but showing consistent signs of improvement. Others might be performing above the average, but starting to slip. Is this something to be alarmed about? Which course should be prioritised? It is essential that we are able to capture changes in student satisfaction over time.

Devaluation of free text

Many current practices in the institutional reporting of SETs also lead to a situation where the free text component of SETs is given less value. A choice is often made to consider numeric data that can be easily analysed (using e.g. averages, standard deviations, *p*-values) rather than what *ought* to be analysed (e.g. sentiment, thematic clusters in response formats, correlation of satisfaction to a chosen curriculum pathway).

This focus upon numerical responses and the associated devaluation of more complex data is unfortunate. For example, free text responses could often reveal a wide variety of essential contextual information that help us to understand average scores. Thus, what comment was left by the lone student who gave the course a 1? Perhaps they are deaf and complaining about a lack of organisational support, or maybe they felt that they were continually harassed by the lecturer. Each scenario would require a markedly different response from a manager. A well implemented evaluation framework allows for numeric scores to be linked to individual free text responses, which

provides academic staff with additional context to understand the reasons behind individual responses. But this is by no means always the case.

Contextualisation to organisational unit

It is common for SETs to be compared across inappropriately large organisational units in a criterion referencing scenario (Abrami, 2001). Performance metrics are often defined at a university level, but this fails to capture the manner in which different organisational units might be achieving markedly different distributions of SET scores (Aleamoni, 1999). This means that an average value that is deemed 'underperforming' in one organisational context might be considered very much on par in another one. For example, suppose that a university examined all of its SET data for a 5 year period, determining that the average score across the entire organisation for this period was 3.9. It would then be very easy for that organisation to declare some minimal set of thresholds below which a course (or academic staff member) would be determined as 'underperforming' and another one above which they would be declared 'performing'. How would such a scenario be likely to play out?

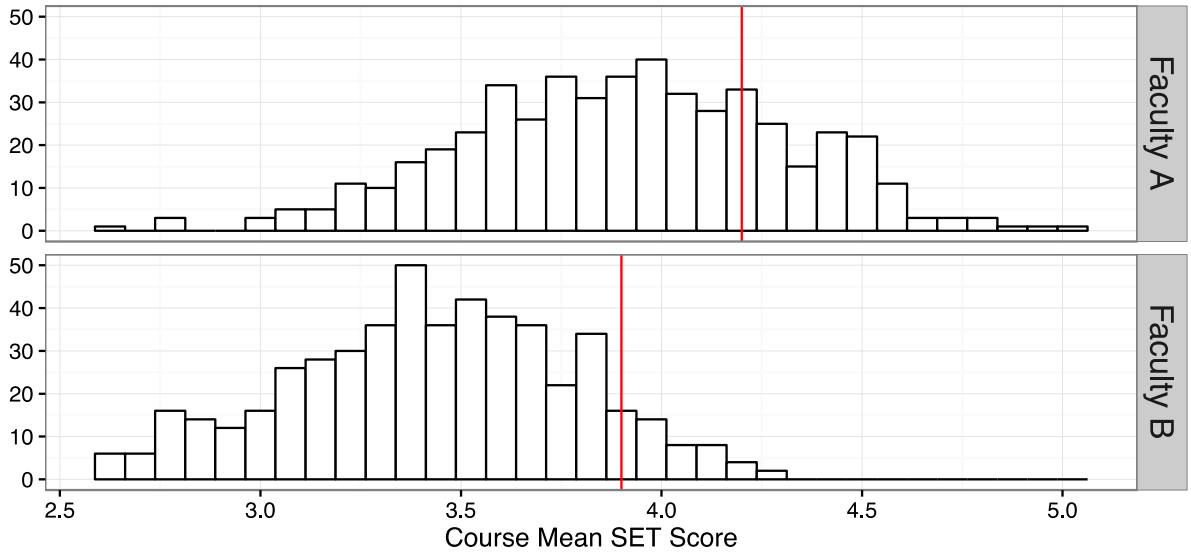

Figure 2. A demonstration that two average scores occurring for courses in two different faculties could imply very different teaching performance, even to the extent that the highlighted course in Faculty B is likely to be performing better than the one in Faculty A despite a lower average value. (Simulation sizes 500, 300).

Let us refine this scenario with reference to the data plotted in Figure 2. Here we see two artificially generated distributions of average scores for two imaginary faculties. If both of these distributions were obtained from the same university then it would be highly problematic to use a university-wide average value as a measure of performance. Indeed, the bulk of the units in Faculty B are likely to be deemed as 'underperforming' in a comparison across the whole university. But are they? Considering two specific courses which obtained the averages depicted by the vertical (red) lines shows that the line in Faculty B occurs at a lower value than that for Faculty A, but it could be considered to be a top performer in the context of its faculty. What is the valid unit of comparison?

This issue is similar to the defendant's fallacy (Low Choy and Wilson, 2009) which occurs when poor results are effectively 'diluted', and hence obscured, by pooling them with better results. The opposite may also occur: the prosecutor's fallacy, where poor results may be emphasized when corralled within a small subset of students, and hence overstated.

Supporters of norm referencing suggest that contextual factors such as organisational unit should be incorporated into the analysis of SET data in order to generate valid comparisons, preferably across many different organisational levels, and/or teaching contexts (see Abrami (2001) for a clear discussion of the merits and pitfall of both norm and criterion referenced reporting for SETs). The debate continues, but many criteria referencing systems are defined with respect to a norm (i.e. in the definition of an absolute standard of teaching performance using SETs it is common to analyse existing data). Therefore, we consider it essential that techniques be developed that can be used to compare student satisfaction within different institutional contexts.

Likert items vs Likert scales

There is a quantitative distinction to be drawn between individual Likert items (i.e. specific questions in a basic survey) and a systematically developed Likert scale. Constructing a scale requires an extensive and careful approach, which includes the selection, analysis, and ongoing refinement, of a set of questions that are deemed representative of an underlying set of latent psychological characteristics (see e.g. DeVillis (2012) for an intuitive introduction, and Worthington and Whittaker (2006) for a set of recommendations as to best practice). There is no guarantee that the latent variables correspond precisely to what is first hypothesised, and it can be very difficult to extract the underlying meaning of a set of items.

Educational testing is not always carried out by those trained in fundamental skills of constructing scales: calibration, validation and testing for reliability. Universities often fail to test the validity of a scale against their cohort, even if they are adopting a well understood construct that is generally considered valid (Spooren, Mortelmans, and Denekens, 2007). Some SET scales have been verified (Abrami, D'Appolonia, and Rosenfield, 2007; d'Appolonia and Abrami, 1997; Marsh, 2007) but few replication studies exist. One example is provided by Rindermann and Schofield (2001), who demonstrated the validity and reliability of their instrument across six traditional and technical German universities. Other notable exceptions arise when scales are applied across cultural contexts, for example, see Mittal, Gera, and Batra (2015) who perform a replication study of the scale reported in Shevlin et al. (2000) in India, and Marsh et al. (1997) which studies a Chinese version of the Students' Evaluations of Educational Quality Instrument (SEEQ) (Marsh, 1982). Not enough of these replication studies have been attempted, which means that there are few reasons to believe that a scale constructed within the context of one university will be valid in another. Furthermore, as Spooren, Brockx, and Mortelmans (2013) have pointed out, even verified scales should be re-verified as the student base and teaching practices of an institution evolve.

Even more problematic, it is common for universities to take a series of questions, or even a single item, that decision makers feel will provide insights about teaching quality, and then use them *as if they were* a verified scale (Spooren, Mortelmans, and Denekens, 2007). There is no guarantee at all that results gathered in this manner will translate to another institutional context, or even hold validity in the context where they are being used.

Ordinal data

Even beyond these issues, a fundamental one of analysis presents: the use of average values can be highly problematic for Likert items. As Jamieson (2004, p1218) succinctly states: "the average of fair and good is not fair-and-a-half". Likert items are typically recorded on an ordinal scale, which means that the difference between 1 and 2 may be substantially larger than the difference between a 3 and 4, despite their numerical equivalence. Even solving this problem, what of the student who selected 'neutral'? The midpoint of a Likert item can mean more than one thing, such as neutral, a mix of positive and negative, unsure/don't know, don't want to answer/commit. At this point we see that using an average value in a high stakes performance framework can become highly problematic. It can encourage a disconnect between numbers and their underlying meaning, leading to a significant misuse if naive interpretations are adopted. For example, an assumption that ratios are preserved can be problematic and must be tested. If distance is not preserved then statistical measurements like average, standard deviation and ANOVA become questionable. Does the student who 'strongly agrees' that a unit has helped them to learn have twice the agreement of a student who merely 'agrees'? While interval scales are normally assumed when analysing SETs, this is an assumption that needs to be tested on the data. Depending on the wording of the item, and upon the various ways in which different student cohorts might interpret such questions, we can anticipate that Likert items will sometimes not be well represented as interval. This is a point at which we would need to make use of more sophisticated methodologies.

Furthermore, while statistical texts routinely declare that the median and the mode should be used for ordinal data (Blaikie, 2003), simple distance based metrics are often reported for SETs without this important clarifying information. We might ask why there is such a preponderance of work that seeks to simplify its analysis inappropriately; the multinomial distribution is the model of choice for ordinal data in statistics (Gelman et al., 2013). This further suggests that a change in methodology is appropriate for the SET field.

Multiplicity and the replication crisis

Even a brief examination of the literature that attempts to correlate SETs to teaching performance reveals that the field contains many contradictory findings (Aleamoni, 1999; Spooren, Brockx, and Mortelmans, 2013). There is, no doubt, a large amount of institutional variability in how SETs are used, which will cause many genuinely contradictory findings. However, a second explanation is likely to be possible for at least some of these results; hypothesis testing is a fraught enterprise. A significant p-value (e.g. $p<0.05$) is by no means a guarantee of a real effect and often prone to mis-interpretation (Greenland et al., 2016). This makes it entirely possible that many results declaring as 'significant' some correlation of SET responses with underlying bias, low response rates, grade related answer patterns etc. are likely to be false positives (Gelman and Loken, 2014; Nuzzo, 2014).

This problem is often referred to as multiplicity, and has led to what is now termed the *replication crisis* in a number of fields. Its origins lie in the many different ways in which hypotheses can be selected - a phenomenon often referred to as 'p-hacking' or 'researcher degrees of freedom'. However, as is compellingly argued by Gelman and Loken (2014) such outcomes need not imply that researchers are actively performing multiple illegitimate tests. They are collecting data about complex social scenarios; in each case it is possible to collect and then analyse this data using well thought out and theoretically plausible methods, and yet for most real datasets many other choices could also have been made. This can have the result of a `significant' value that is due more to chance rather than a real underlying phenomenon. Similarly, many studies may not provide

significant *p*-values, but still be informative when analysed in different ways (Western and Jackman, 1994). The debate on these problems with *p*-values has simmered for decades in the field of education (Fidler and Cumming, 2005; Myer, 1964; Thompson, 1996), but is yet to affect core practice. The time is now ripe for action; a recent public statement by The American Statistical Association points to the misuse of *p*-values in many disciplines (Wasserstein and Lazar, 2016); clearly it is inadvisable to ignore such advice. We contend that problems with replicability are likely to be rife in the SET literature, and could be the source of the many contradictory results that have arisen in the field. New methods are required to move forwards.

A new approach using contemporary statistical techniques

A number of people have proposed methodologies for avoiding some of the pitfalls raised in this section. For example, Neumann (2000) has suggested that an approach using rating interpretation guides (RIGs) could take into account different teaching contexts, also emphasising that a range of SET values rather than an average score was most effective. Similarly, Abrami (2001) proposed a set of detailed criteria for the interpretation of statistical analyses of SET scores (in the context of hypothesis testing).

While we consider such approaches worthy, the problem is not so much with the interpretation of the analysis, as the with analysis itself. Contemporary statistical methods for analysing SET data would avoid many of the pitfalls that we have discussed above, and in what follows we will demonstrate one way in which this might be achieved. Many other approaches are also possible.

In what follows we will explore some of the issues that we have raised above with reference to the evaluation framework adopted at QUT. A more appropriate statistical methodology for analysing this data based upon a hierarchical Bayesian model will then be introduced. Our method is a simple first step, and is general enough that it can be refined and extended to account for other contextual factors. However, the resulting model is complex, and building a framework that will assist both decision makers and academic staff with sense-making is essential (Kirschner, Buckingham-Shum, and Carr, 2012).

Case study: QUTs Reframe methodology

Here we consider one example of an evolving SET methodology for one university, Queensland University of Technology (QUT). In 2011 the university received strong feedback from academic staff that the online SETs in use from 2007-2011 took into account neither the complex and changing nature of teaching, nor the diversity of contextual environments in which they were deployed. This implied that QUTs SET regime lacked reliability and validity for a modern context that increasingly used online, blended, and other flexible modes of delivery. Furthermore, the purpose of data collection was questioned; was the focus on accountability or on the improvement of learning and teaching? This prompted the launch of Reframe, a five-year project, aiming to give academic staff agency and so bring about widespread organisational change through an evaluation framework.

The Reframe project consisted of a purposeful literature review, a national scan of university practice, and a design-led process to engage with internal and external stakeholders through committee meetings, working groups, campus roadshows, interviews and focus groups with students and academic staff (Alderman and Melanie, 2012; Alderman, Towers, and Bannah, 2012).

Survey methodology

As a result, in 2013 Reframe delivered three new online surveys. It is important to note that they were predominantly developed through this process of stakeholder engagement rather than following the path of a validated construct. The surveys deployed were refined through pilot testing of several instruments with 100 academic staff and 6,600 students, along with a series of focus groups.

The methodology involves delivering a *Pulse* Survey early in the semester, straddling the date at which students could elect to alter enrolment choices. This provides early actionable feedback to instructors, a process which is followed by an *Insight* Survey, deployed late in the semester from the last teaching week across the complete examination period. An *Exit* Survey is also sent weekly (between weeks 2-12) to every student who has withdrawn from a course. Finally, academic staff engaged in teaching into a course are also invited to provide feedback on students' perceived engagement in that course. By the end of 2016, 1.7 million lines of data had been recorded across the university using this methodology.

As a formative tool used to improve teaching practice, the Reframe approach was considered useful by many staff. For example, the early Pulse data was often used by both managers and teaching teams to reveal ways in which course offerings could be improved during the teaching period. However, in 2016, new institutional directions led to the creation of a performance metric based upon the average value of the Q3 item in the Insight survey ("Overall, I am satisfied with this unit"). The study discussed in this paper arose from an attempt to explore possibilities of using modern statistical methods to achieve more nuanced measures of teaching performance within this changing institutional context.

In what follows we will make use of the Reframe dataset to demonstrate that a number of the concerns we have discussed in Section 2 do indeed arise in what could be considered a standard institutional dataset.

Data

In Figure 3 we see the distribution of average overall satisfaction for all major faculties (not identified). We see that at this organisation, the pattern of satisfaction is skewed towards higher ratings. While the distributions are largely similar, units in some Faculties do appear to be achieving higher and/or more consistent satisfaction ratings on average, with e.g. Faculty C exhibiting a tight spread in values, and Faculty B achieving a higher proportion of perfect scores than all other faculties (with the possible exception of Faculty D). This difference in patterns gave us reason to be cautious about criterion referencing - a performance metric that was not contextualised to the organisational unit in which it occurred would be prone to misinterpretation, e.g. inappropriate classification of academic staff with lower averages as underperforming, even though they might be achieving far better satisfaction scores than peers in their faculty. Drilling down to the level of a school adds even more complexity, with the distribution of satisfaction scores obtained by the four schools in Faculty D illustrated in Figure 4.

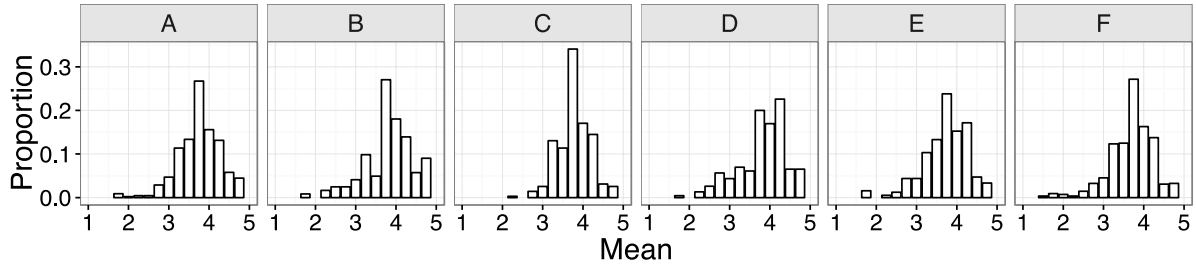

Figure 3. The distribution of average scores obtained for the 6 different Faculties at QUT (not labelled) in 2016.

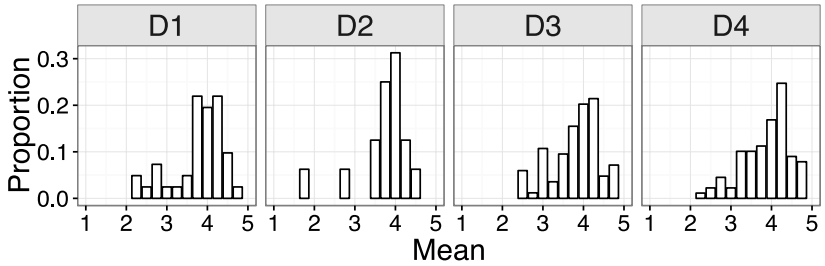

Figure 4. The distribution of average scores obtained for each of the four different Schools in Faculty D in 2016.

It is also very easy to demonstrate that these averages are hiding a large amount of extra detail. For example, in Figures 5-7 we have depicted the distribution of scores obtained for 6 different courses in different faculties. While each figure depicts courses with the same average, together they exhibit all of the signatures discussed in Section 2. It is worth noting that as the average score approaches the extremes of 1 or 5 there are less ways (albeit still numerous) in which ratings can be combined and still obtain the same score.

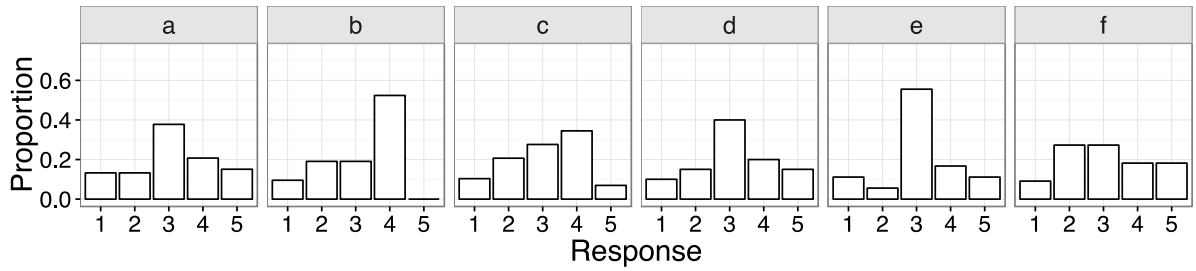

Figure 5. Courses which markedly different distribution, yet share the same average score for overall satisfaction of 3.1 for Faculty E in 2015 Semester 1. (Sample sizes: 53, 21, 29, 20, 18, 11.)

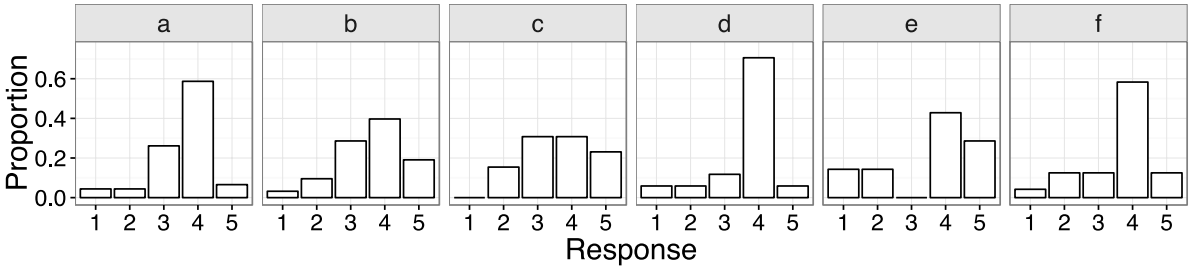

Figure 6. Courses which have markedly different distribution, yet the same average satisfaction of 3.6 for Faculty C in 2015 Semester 1. (Sample sizes: 46, 63, 13, 17, 7, 24.)

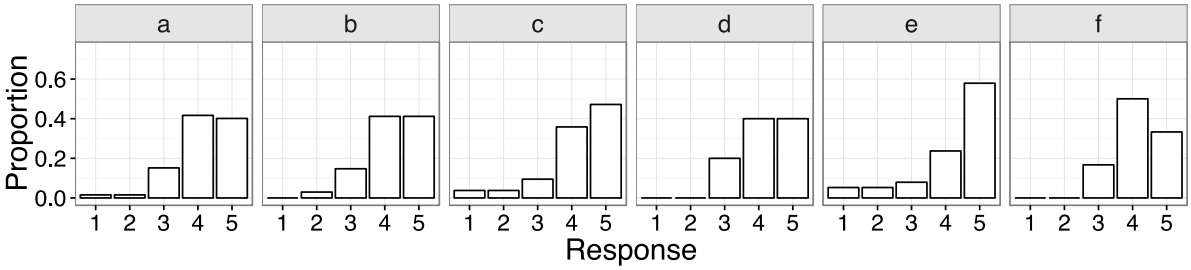

Figure 7. Course which have a markedly different distribution, yet the same average score for overall satisfaction of 4.2 for Faculty B in 2015 Semester 1. (Sample sizes: 132, 34, 53, 5, 38, 6.)

Even more structure emerges if we start to consider the way in which courses change in time: from Pulse to Insight survey in one teaching period (Figure 8), to the change in Insight scores from one year to another (Figure 9). This is important information for decision makers to consider when allocating resources to courses, or in prioritising interventions. It may also be highly indicative of an improving/worsening performance over time, and so could factor into performative frameworks if it could be reported upon in a sensible way. Rather than following Abrami (2001) and aggregating information over a number of years, we would prefer to be able to capture information about how ratings for a specific teaching team or course changes over time.

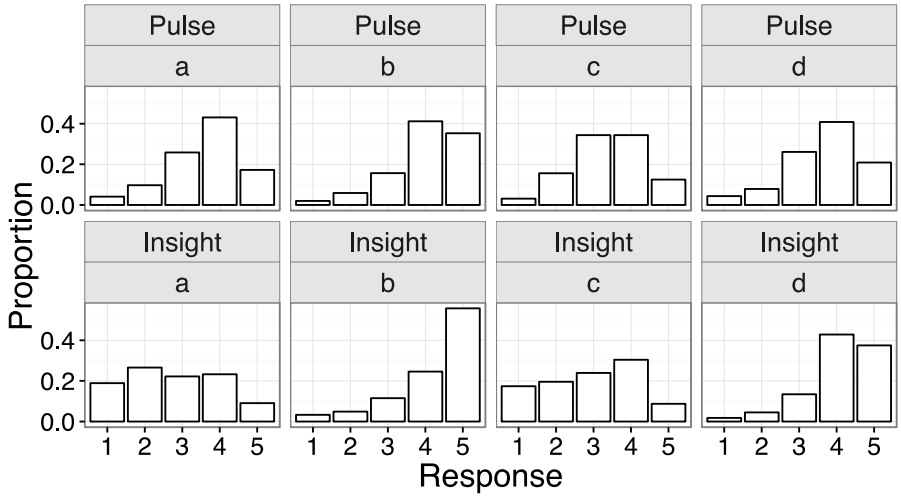

Figure 8. The change in distribution of Q3 values, from Pulse to Insight, in satisfaction, for four courses. Sample sizes are 267 51, 32, 115 for Pulse and 365, 61, 46, 112 for Insight.

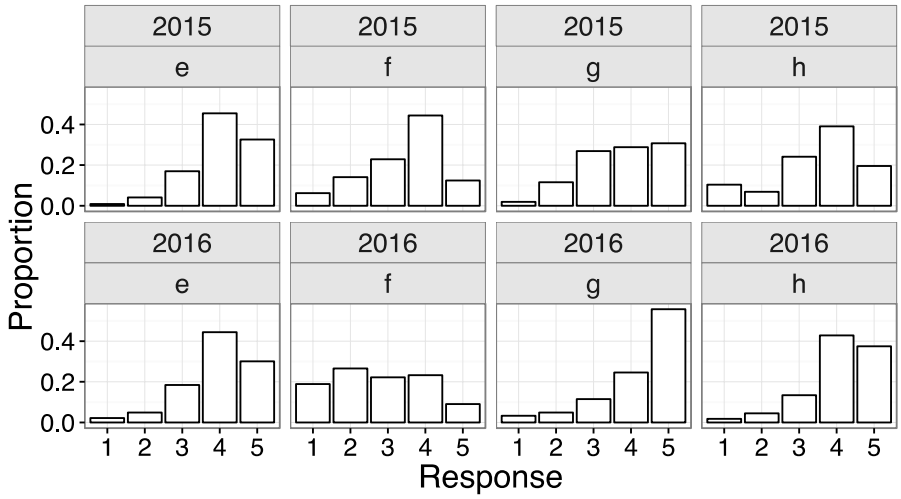

Figure 9. The change in distribution of Q3 values that can occur from year to year for a series of courses. Sample sizes are 365, 306, 52, 87 for 2015 and 369, 365, 61, 112 for 2016.

Moving forwards with Hierarchical Bayesian techniques

We decided to analyse the Reframe dataset using contemporary statistical techniques. Multinomial models are the common modern method for dealing with ordinal data, and Bayesian approaches are well known for circumventing problems of multiplicity, so we decided to pursue a model with these characteristics. Similarly, as SET data arises from a hierarchical university structure (i.e. where students enrol in courses, which are offered by Schools, which belong to a Faculty), we decided that a hierarchical model would provide more nuanced estimates about expected SET responses, and measures of deviation from them, for each level of the organisation.

Rather than focussing upon an average score for a given course, the model we present here describes the distribution of values of the score, within a given organisational unit, across the range of its Likert scale responses {1,2,3,4,5} in this institutional context, but other ranges could be similarly modelled). This distribution can be compared with the distribution of scores obtained in some containing institutional context (e.g. a school or faculty). Thus, our model is designed to be norm referenced.

Our model is also designed to allow a comparison of the way in which a distribution of SET scores is changing in time (e.g. from year to year). This enables an understanding of how student evaluations are changing within a particular context. In this section we will present the basic model that was implemented at QUT for the reframe dataset, although it is important to be aware that many other models are possible, and depending upon the questions to be answered a different hierarchical structure may be necessary in a new institutional context. In what follows we will gradually introduce the ideas that lie behind the model we adopted. We note that while the discussion of the full model will be quite technical, it can be skipped while still gaining a feel for the approach adopted here.

Bayesian models: the core idea

The basic idea of Bayesian modelling (Gelman et al., 2013) revolves around a very simple intuition; gaining further knowledge about a system enables us to update our beliefs about events that are likely to occur in it. We explicitly acknowledge our starting beliefs as a hypothesis *H* and some evidence *E* that we have already gathered. Prior information can be used to evaluate the plausibility of a range of hypotheses, via a *prior*, *p(H)*. Then we use a standard statistical sampling model to describe how likely the evidence is to occur given our hypothesis, the *likelihood*, *P(E|H)*. Given this information, we can write a conditional probability, termed the *posterior probability*, for the plausibility of the hypothesis given the evidence observed

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \qquad (1)$$

where *P(E)* integrates the likelihood over all plausible hypotheses $P(E) = \int P(E|H)P(H)dH$, and essentially serves as a normalising term to ensure the total probability, of all possibilities, amounts to one. Equation 1 is known as *Bayes Theorum.* While it is remarkably simple, its interpretation has attracted a large amount of both controversy and confusion. Much of this revolves around the definition and use of the priors. We will not delve into these murky waters here (although see Low Choy (2012) and Myer (1964) for good introductions). Instead, we will consider why a Bayesian approach might be preferred when attempting to understand large SET datasets.

Why a Bayesian model?

Bayes' Theorem allows us to reformulate our thinking. It supersedes asking how likely our data is, when a particular hypothesis holds true, *P(E|H)*, which is the standard approach followed by a regime based upon hypothesis testing. Instead Bayes theorem allows us to consider the probability that a hypothesis is correct (i.e. its plausibility) given the data and hence evidence we obtained, *P(E|H)*. This allows those using Bayesian statistics to express their confidence for a particular parameter being in any particular range rather than setting an arbitrary cut off for significance and then testing only one available hypothesis. A frequentist approach based upon hypothesis testing would seek to answer questions such as: "Is the SET score achieved by this course significantly different from the average value achieved by the school?" and "What range of average SET scores makes our data most likely?" In contrast, a Bayesian approach allows us to ask questions like: "Based on the data we have observed, what is the plausible amount of difference between the profile of this course compared to that of the school?" This contrast means that a Bayesian approach can start to correct for the problem of multiplicity that is likely to lie behind the many false positives that we believe beset the SET literature.

A two level hierarchical Bayesian model

A hierarchical, or multilevel, model takes into account the structure of a dataset, to describe its pattern of variation as well as features like the average. Hierarchical Bayesian models make use of Bayes Theorem, noting that hypotheses *H* can be specified in terms of model *parameters*, which are supported by evidence *E in the data:*

$$P(params|data) \propto P(data|params)P(params) \qquad (2)$$

We can incorporate relationships among variables in the data using this methodology. If the sampling distribution of the data, given the parameters, also depends on explanatory variables *X* then (2) expands to:

$$P(params|data, X) \propto P(data|params, X)P(X|params)P(params) \quad (3)$$

The basic Bayesian model can also be expanded to allow parameters to depend on `hyper-parameters', which are parameters of the prior distribution:

$$P(all\ params|data) \propto P(data|params)P(params|hyperparams)P(hyperparams) \quad (4)$$

Let us consider a simplified scenario for the sake of illustration. We will seek a model that explains student responses about overall satisfaction. In this model we will assume that satisfaction depends upon the course in which the student is enrolled, and the school in which that course is offered. Thus, the *data* comprise the responses from each student *i* in course *c* in school *s* which we denote as $R_{i,c,s}$. Then we can describe the average satisfaction level that arises from considering all students in the course using the parameter $\nu_{c,s}$. Similarly, the average satisfaction level obtained by considering students in all courses in the school is captured by the parameter ω_s. Applying hierarchy to these parameters (Equation 4) we can construct a hierarchical model using the following decomposition:

$$P(\nu, \omega|R) \propto P(R|\nu, \omega)P(\nu, \omega) \quad (5)$$
$$\propto P(R|\nu)P(\nu|\omega)P(\ \omega) \quad (6)$$

Normal model

In the above basic model, patterns in the data, and our uncertainty in each of the relevant parameters can be straightforwardly described by normal distributions:

$$\text{student responses: } R_{i,c,s} \sim N(\nu_{c,s}, \rho_c^2)$$
$$\text{students in course: } \nu_{c,s} \sim N(\omega_s, \chi_s^2)$$
$$\text{courses in school: } \omega_s \sim N(\xi, \zeta^2)$$

The terms ω_s, ξ are called *hyper-parameters* and have hyperprior distributions represented by their standard deviations χ_s, ζ(the standard deviation, ρ_c of the parameter describing student responses is given by the data). This model describes the probability of a particular level of satisfaction arising from a student in a given course, centred around the average satisfaction in the course, and then school. This approach makes use of information that has already been learned (which in this case is the average satisfaction for that course) to determine the probability of a given distribution of student responses. In summary, this hierarchical approach breaks the probability model up into three levels, and considers evidence gained from one level in the construction of the model for the next. Inferences can then be made using the parameters ν, and hyper-parameters ω_s, ξ to estimate the pattern of responses for courses, schools, and with an extension of the model, the university as a whole.

Previous Bayesian models of SETs

This form of analysis is by now prevalent in other fields, but is surprisingly rare in the SET field. However, some people have drawn attention to the need to modernise our approach to the

evaluation of SETs. In responding to Abrami (2001), Theall (2001) notes that McKeachie has advocated the use of Bayesian approaches over hypothesis testing, but few implementations of this suggestion exist. One example is provided by Wetzstein, Broder, and Wilson (1984) who demonstrate an example methodology for determining the difference in feedback obtained for a graduate student and a professor (i.e. a single level analysis which misses many of the structural features common to SET data). Huang and Wang (2014) constructed a set of two-level (student/class) hierarchical Bayesian Item Response Theory models that considered whether student scores of 'overall teaching effectiveness' were predicted by the gender of the instructor as a level 2 covariate. They found no support for this hypothesis, but were able to demonstrate that the extra structure provided by the hierarchical model was necessary in drawing this conclusion. However, it is worth noting that both of these models are much less complex than the one presented in this section. On a slightly different note, Bayesian models have also been constructed where SET scores are used as predictive variables. For example, Galbraith, Merrill, and Kline (2012) utilise a Bayesian data reduction algorithm to classify student learning using variables that include SET scores.

We see that the technique is not entirely new, but that it has yet to enter into any form of systemic organisational usage for the analysis of SETs, perhaps through a lack of time, familiarity, or expertise among those who have access to university wide datasets. In what follows we will present the general technique that we have used to construct a full Hierarchical Bayesian model over the 1.7 million responses covered by the Reframe dataset. It is hoped that the techniques introduced here will encourage the wider usage of a standard contemporary statistical method which enables a far more nuanced exploration of SET data in a range of institutional contexts.

Multinomial model: The full Reframe model

In constructing the full Reframe model, we will reconsider the Normal distribution used in the basic model, and cease to require an average SET value as a proxy for the performance of a course. Instead, we will now make use of a multinomial distribution to consider the proportions of different scores for a specific school; under the Reframe regimen which gives 5 possible responses on satisfaction: {1,2,3,4,5} encoding levels from very dissatisfied to very satisfied. We will also expand the number of variables that we use to explain satisfaction. We will then seek to establish whether the performance of a course is significantly better or worse than the average for its school, considering the whole range of possible responses.

We are also interested in changes in response over time. We will aim to model how the satisfaction scores obtained by a course change: from Pulse to Insight in one semester, as well as from year to year. These comparisons will again be contextualised with reference to the school.

Hierarchy of structure

Keeping these two requirements in mind, we need to understand the way in which the proportion of satisfaction responses obtained by a course compares with the distribution across all courses in its host school. We represent the proportion of scores in a course (which is {1,2,3,4,5} in this case) using $\pi = (\pi_1, \pi_2, \dots, \pi_5)$, and our uncertainty about that proportion using $P(\pi)$.

The data for our model comes from student responses to the satisfaction item, denoted by $X_{r,m,y,t,f,s}$, where the subscripts indicate how this response relates to other parameters specific to the Reframe model:

r=1,…,R, possible responses for satisfaction (*R*=5)

m=1,…,M, the semesters in a year (*M*=2 in this model)

y=1,…,Y, the year (this study takes 4 years, i.e. *Y*=4)

t=1,…,T, the survey types (in this case *T*=2: Pulse and Insight)

f=1,…F, the faculties (at QUT this was 6)

$s=1,\ldots S_f$, the number of schools in a specific faculty

(8)

Modelling probability of Likert-scale responses

We model our data $X_{r,m,y,t,f,s}$ as being sampled from a Multinomial distribution (which allows for any number of possible categorical outcomes)

$$X_{r,m,y,t,f,s} \sim Multinomial(\pi_{r,m,y,t,f,s}, n_{r,m,y,t,f,s}) \quad (9)$$

where $n_{r,m,y,t,f,s}$ is the number of completed surveys in a subgroup of students (e.g. the school) as defined by subscripts in Equation 8, and $\pi_{m,y,t,f,s} = (\pi_{r,m,y,t,f,s}; r = 1 \ldots R)$ is their distribution across the *R* possible Likert-scale responses. This requires π to be positive and normalised (i.e. all probabilities must sum to 1). We apply the following transformation:

$$\pi_{r,m,y,t,f,s} = \frac{\theta_{r,m,y,t,f,s}}{\sum_{r=1}^{R} \theta_{r,m,y,t,f,s}} \text{ where } \theta_{r,m,y,t,f,s} = e^{\psi_{r,m,y,t,f,s}} \quad (10)$$

which ensures this, re-centres our model, and allows us to model our uncertainty in the $\psi_{r,m,y,t,f,s}$ parameters using a Normal distribution (see Equation 12 below). The support of the Normal distribution is $\psi \in \Re$, which when exponentiated becomes $\theta = e^{\psi} \in \Re^{+} \cup \{0\}$.

Full model

We can now write a more complex hierarchical model, starting with a form similar to Equation 7, but extending with extra terms of interest in the hierarchy:

$$P(\psi, \phi, \eta, \gamma, \mu, \xi | X) \propto P(X|\psi)P(\psi|\phi)P(\phi|\eta)P(\eta|\gamma)P(\gamma|\mu)P(\mu|\xi)P(\xi) \quad (11)$$

where each of our parameters are assumed to follow a normal distribution, centred at an average effect (on the transformed scale) relevant to that level of the hierarchy:

$$\begin{aligned}
\text{school average:} \quad & \psi_{r,m,y,t,f,s} \sim N(\phi_{r,m,y,t,f}, \sigma^2_{r,m,y,t,f,s}) \\
\text{faculty average:} \quad & \phi_{r,m,y,t,f} \sim N(\eta_{r,m,y,t}, \tau^2_{r,m,y,t,f}) \\
\text{survey type average:} \quad & \eta_{r,m,y,t} \sim N(\gamma_{r,m,y}, \alpha^2_{r,m,y,t}) \\
\text{cohort average:} \quad & \gamma_{r,m,y} \sim N(\mu_{r,m}, \kappa^2_{r,m,y}) \\
\text{semester average:} \quad & \mu_{r,m} \sim N(\xi_r, \lambda^2_{r,m})
\end{aligned} \quad (12)$$

Finally, the average for a semester centres on a global average, which given no previous information, is allocated a vague prior, Γ_r:

$$\xi_r = N(0, \Gamma_r). \qquad (13)$$

This assumption reflects a lack of knowledge about variability of SET scores, a modelling assumption that can be assessed via sensitivity analysis (Gelman et al., 2013) but could be relaxed for other analyses according to the needs of an institution.

The model constructed in this way assumes that student responses cluster around their school average, which across schools centres on a faculty average. Then the faculty average centres on the average for that survey type, then cohort (i.e. degree program) and then semester. Note that the score for a specific course is not included in the model; it will reappear in Section 5, where we demonstrate how the model compares the SET responses received by a course with the average values at the school, faculty etc. levels. Thus, this model helps us to understand the *expected* responses to a SET at each level in the hierarchy constructed. These will then be compared with the values that a course actually obtains to extract information about how it deviates from that expected value. Note also that many other hierarchical orderings could have been constructed. It depends upon the requirements of the analysis, and what organisational units it makes sense to compare over multiple data collection points.

Once responses are suitably transformed to a normal distribution, we are provided with an elegant and parsimonious way of describing a nested hierarchy of average effects, at increasing scales of aggregation (Gelman, Hill, and Yajima, 2012). All of the random effects $(\sigma, \tau, \alpha, \kappa, \lambda)$ are assigned independent zero-truncated normal priors (Gelman et al., 2013), which for numerical reasons are truncated just above zero in our implementation (which used OpenBUGS (Sturz, Ligges, and Gelmann 2005).

Modelling changes between variables

Having constructed a model, we can start to explore some of the questions raised above (in Section 2), using a contrast that quantifies the difference in proportion of scores, between two years:

$$\delta_{r,m,y,t,f,s} = \pi_{r,m,y,t,f,s} - \pi_{r,m,y-1,t,f,s}. \qquad (14)$$

Or for the same semester, a comparison between the Pulse and Insight scores obtained for a course:

$$\gamma_{r,m,y,t,f,s} = \pi_{r,m,y,t,f,s} - \pi_{r,m,y,t-1,f,s}. \qquad (15)$$

Many different questions are possible, depending upon what is considered organisationally important, and the data that is available.

Using the full Reframe model

The Bayesian Hierarchical Model produces posterior estimates of the average proportion of each satisfaction score given a set of conditions, which allows us to explore both the way in which different organisational conditions affect responses, as well as how these responses are changing in time. In this section we will explore some of the ways in which this model can be used to help an institution understand changing patterns of student satisfaction and how they might be contextualised to various organisational units.

The models constructed take the posterior predictions for a specific level of the hierarchy and enable a comparison of the course of interest to the distribution generated by the posterior.

Thus, it is at this point in the model that the proportion of students responses in a course (i.e. the proportion of Likert scale responses {1,2,3,4,5} denoting very dissatisfied through to highly satisfied that are obtained in a specific subject) are compared to the likely distribution of responses that are obtained for the organisational unit.

Comparing the performance of a course to its host school

Figure 10 provides information about the $P(\pi|X)$ probability distributions for two schools in our dataset (the curve), along with the proportions of responses for a specific course in each school (the straight line). The area under the curve to the left of the line represents the probability that the school's mean proportion of responses for that score is less than the proportions for that specific course. The area to the right of the line gives the probability that it is greater. For example, considering a satisfaction response of 5 (i.e. 'highly satisfied') in Figure 10(a), it can be seen there is a very high probability that the school mean is less than the course's results. This can be seen by the fact that the line (representing the proportion of 5 scores obtained by the specific course) is to the right of the density curve (which represents the posterior distribution estimated by the model). This position of the line suggests that the course is being rated with noticeably more 5 responses than comparable other courses its host school. This would suggest that the course has significantly more highly satisfied students than is the norm for that school. Looking through the rest of the possible responses shows us that overall it has a lower proportion of 1, 2 and 3 scores than the school, and more 4 and 5 responses than could be expected. It appears that the students in this course are on the whole more satisfied than is usual for this school.

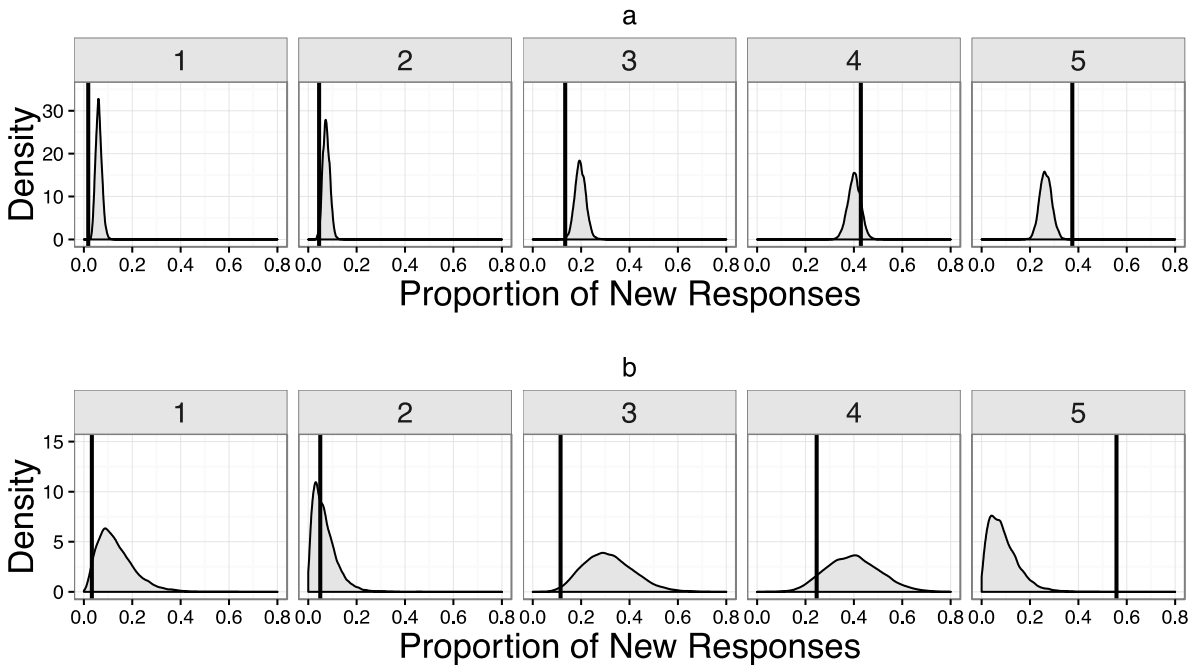

Figure 10. Posterior distributions for the proportions of the multinomial responses (i.e. satisfaction scores {1,2,3,4,5} - listed across the top of the plots) for two schools and two specific courses in those schools for the 2016 cohort. The posterior densities for the school in (a) are narrower than those in (b). The average satisfaction scores for the courses shown in these plots are 4.098 and 3.854 respectively.

Figure 10(b) shows another course that appears to be performing better than its relevant school, appearing to be more likely to achieve a score of 5 (i.e. 'highly satisfied') than other courses in the same organisational unit. However, the greater spread in posterior probabilities suggests that there is a much greater variance in scores being achieved for this school, which gives us reason to be more careful in developing an evaluation strategy for this organisational unit. This difference in school variance suggests our estimate for the likely scores achieved by the school in Figure 10(a) is more certain that that for Figure 10(b). We are less confident about what an expected average of SET scores is, or what distribution of scores is likely to lead to it, for the second school (a factor that becomes particularly important for some other SET scores where the line is overlapping the distribution). This might be due to a number of different factors: from a genuine range in teaching performance; to wildly varying cohorts; or even to the delivery of highly experimental teaching strategies. Extra care must be taken in constructing evaluation metrics in this case. It is not surprising that there will be probabilities with considerable variance when we consider that the model presented here has not included a range of other variables that could be important predictors (e.g. demographics, grade, year of study, learning design etc.). Including more variables could lead to tighter probability distributions, although this may not always be the case.

Evaluating performance over time

Figure 11 shows sample output from a comparison of the way in which the SET responses for two courses have changed from 2015 to 2016, in two different schools. In essence, the distribution δ defined in (14) is represented by the area under the curves. The amount of this distribution which is greater than zero on the *x* axis represents the model's prediction of the probability that the school has increased its proportion of responses for a particular score.

The same figures also compare a specific course's change in proportion of values against the δ distribution of its host school. The area under the distribution to the right of a course's change in proportion is the probability that the course had a lower change in proportion of that response than the school's mean change in proportion.

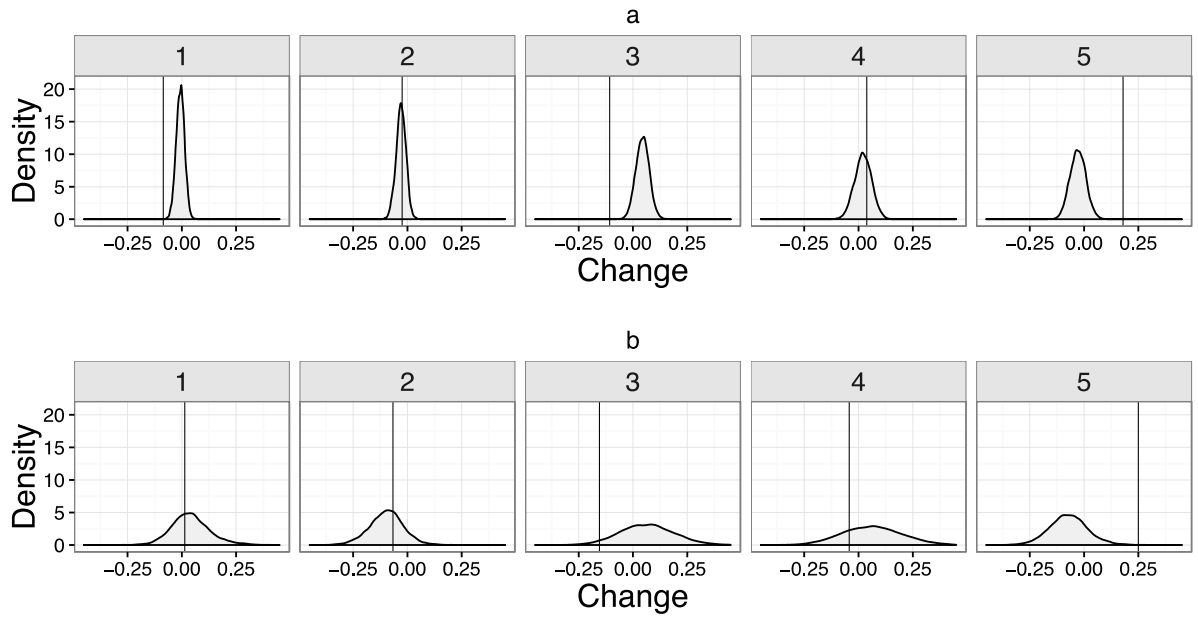

Figure 11. Posterior distributions for the change in school distributions and the change in satisfaction scores for two course both run in 2015 and 2016. Course (a) changed its average score from 3.5 to 4.1. Course (b) changed from 3.8 to 4.2.

For example, Figure 11(a) shows that there are proportionally less responses of 1 occurring for this course in 2016 than occurred in 2015. This is made more interesting by the observation that the school does not appear to have changed at all (as the distribution is still centred on 0). Examining the change in 5 responses for the same course shows that there are now more: the course appears to have improved over time. For this example, the result is in agreement with the mean score, which changed from 3.5 to 4.1 in this period (a clear improvement). However, the new more contextualised report enables us to see far more information about *how* the course has improved. Figure 11(b) shows another example, with a mean increase from 3.8 to 4.2. It can be seen that this increase comes primarily from a decrease in the 3 responses, and a large increase in 5 responses, along with a smaller decrease in the 2 and 4 responses. As was the case for the previous section, we note that the change in the school depicted in Figure 11(a) has much narrower posterior distribution, suggesting the probability of a large change between the two years is much lower than it is for the school in Figure 11(b).

Limitations

The model presented in the previous section could be improved in a number of ways. Due to the vast number of courses present in the data, our estimates were limited to the school level. This avoided problems with both computational time and file size that occurred with attempting to drop to the course level, and helped to alleviate problems with low response rate for some courses. It also avoided the potential overfitting the model to inappropriately small sub-classes in the data. Other institutions may be able to reach a finer level of detail depending upon their SET data.

Another limitation of this model is that it does not consider the differences that can arise with individual students. Where data is available, information on a student's demographics, past academic performance, and historical SET evaluation tendencies could be fed into the model. This would facilitate the analysis of individual variances in how different classes of students might respond to SET items, and in particular a consideration of whether changes in how a course is performing from year to year might be attributable to e.g. a particularly pessimistic cohort.

As with all SET data, we are still left with little information about how non-respondents might differ from those students who responded. However, if the data allows for a model to be constructed at the individual student level, then it may be possible to use models such as this to impute likely responses that would have been given by students who have responded at least once. This form of estimation is made possible by a particular characteristic that we have noticed in the Reframe dataset; most students respond at some point in time throughout their university experience. Thus it seems plausible that models of individual student behaviour could be constructed in the future. We note that this form of study is only made possible in a data collection methodology that enables data custodians to re-identify individual student response patterns. An institution that did not store this data would not be able to construct student level response models. This calls attention to the obvious trade-off between student privacy and accurate modelling; substantial care must be taken by data custodians to ensure that the student feedback is not re-identifiable by e.g. academic staff, or by other user groups who should not have access. While

decisions such as these will be made at the policy level, we consider it essential that data custodians take great care to consider the implications of linking data in their models and to ensure that it is not misused, while championing the need to perform such analyses which can help to improve both the student experience and feedback to academic staff.

Sensemaking with complex statistical models

The model presented in the previous section is not one that can be easily interpreted by those not familiar with statistical models. However, it can be coupled with more intuitive reports to facilitate sense-making for both organisational decision makers (who may need to allocate resources or recognition) and academic staff (who may be seeking to more deeply understand how their teaching is rated by their students). We consider it essential that any sense-making tools be carefully designed to avoid potential abuses. In particular, we would like to avoid misuse through the attribution of meaning to results that are unlikely to be statistically significant.

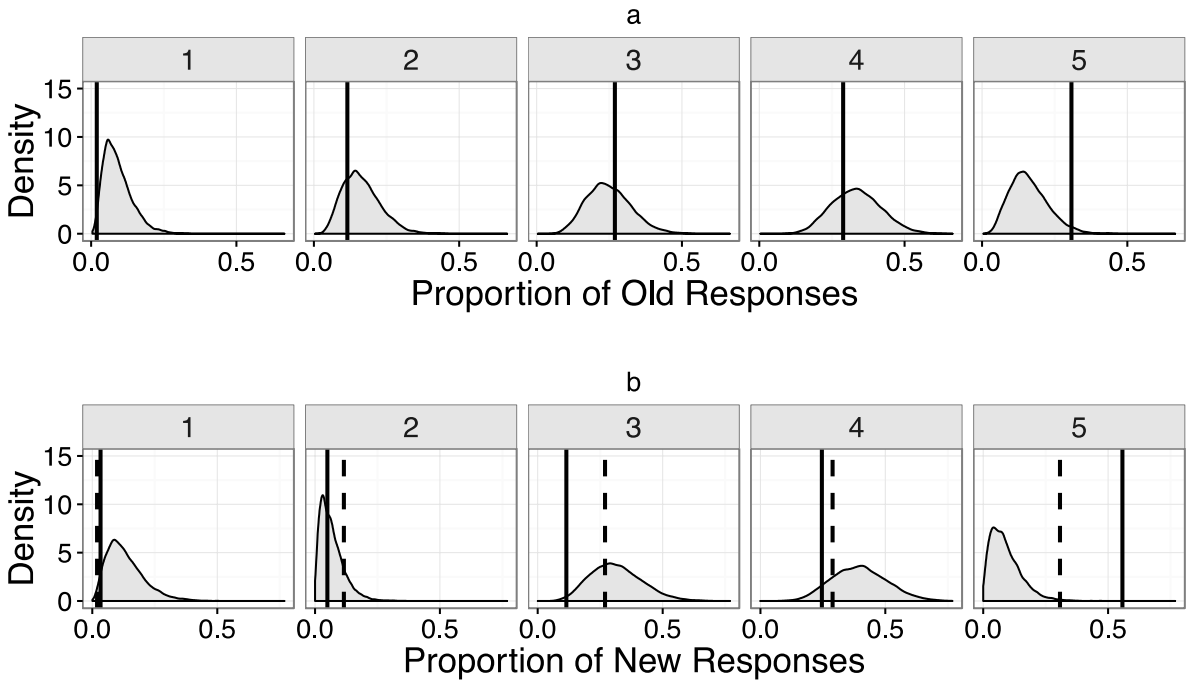

Figure 12. A report following Figure 10 for a new school and course within Faculty A. In this case the posterior distributions for two years are shown. The solid lines represent a course's outcomes in 2016, and the dotted lines the outcomes of 2015. The distance between these lines provides the basis for the *Course Change* row in Figure 14.

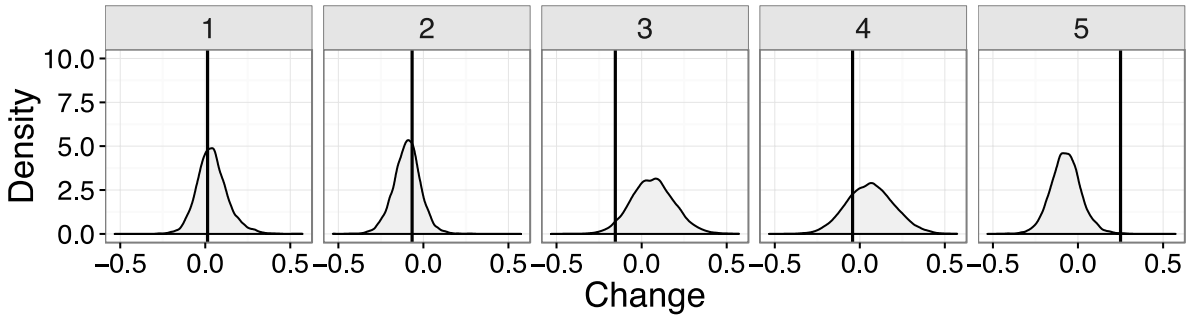

Figure 13. The posterior distribution of the changes in proportions for the scenario illustrated in Figure 12 for 2015-2016. As was the case in Figure 11, the solid lines represent a course's changes between the years. This report is used to generate the *Course Change Compared to School* row in Figure 14.

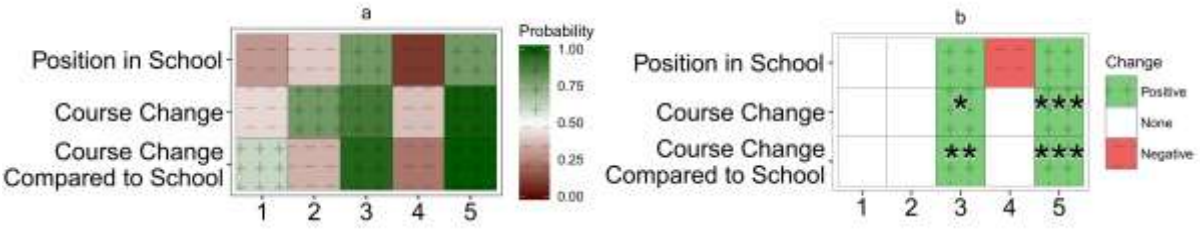

Figure 14. A heatmap can be used as a sense-making device to facilitate a contextualised examination of multiple changes in how students are rating a course over a two year period. Course Change is extracted from the report in Figure 12, and Course Change Compared to School from the report in Figure 13. In (a), these measures are displayed as a gradient of probabilities, while in (b) only changes that cross pre-defined 'significant' levels are reported (*: Significant at the 10% level, **: Significant at the 5% level, ***: Significant at the 1% level). The Position in School cutoff makes use of quartiles, only marking a box as green or red if it is in the highest or lowest quartile respectively.

We shall consider one particular course from the Reframe dataset throughout this section, with all reports generated for semester 1 in 2016. Figure 12 is a reproduction of Figure 10 for our course of interest. We have produced this report for two years, which enables a comparison of how student ratings are changing over the time period for both the course and its host school. In Figure 13 we can see the *changes* in proportions for that same school and course over the same two year period.

The data from Figures 12-13 is used in the generation of Figure 14(a), which is a heatmap created to facilitate more rapid decision making and interpretation of those reports. It is a sensemaking devise, designed to be more intuitively interpretable for those without a strong grasp of statistics, and consists of three panels describing the different components of the course in their organisational context.

Positions in School: This line of the report ranks courses by the proportion of responses that they received for each potential satisfaction score. For responses 1-3, it is desirable to achieve less responses, so we rank from smallest to largest, and conversely for responses 4-5, we rank from the largest proportion of scores to smallest. The colours are then used to rapidly discern where in the list that course was ranked for that Q3 response.

Course Change: This line corresponds to the report shown in Figure. 12. For responses 1-3, a green box is generated if course's proportion is less than the probability of the school's mean proportion, whereas for responses 4-5, a green box is depicted if the course's proportion is greater than the school's mean proportion.

Course Change Compared to School: This line corresponds to the report shown in Figure 13. For responses 1-3, a green box is generated if the course's change in proportion is less than the school's mean change in proportion, whereas for responses 4-5, a green box is depicted if the course's change in proportion is greater than the school's mean change in proportion.

Note that in constructing these reports for the Reframe dataset a decision was made to treat decreases in a satisfaction response of 3 as positive because the mean Q3 response across the entire institution was 3.8 (with a median response of 4), which suggests that a score of 3 is indicative of a course that is being rated lower than average by its students. This could potentially be linked to specific organisational units depending upon the requirements of an institution.

While the representation of Figure 14(a) contains all relevant information, there is a danger that it will be overinterpreted. That is, a continuous scale can lead to small non-significant differences still being represented as bad (i.e. -/red) or good (i.e. +/green), and some people are likely to assume that this difference is meaningful. In order to reduce the risk of this overinterpretation, we recommend a further refinement of the report, only providing colours if the data is suggestive of a practically significant difference (Gelman et al., 2013). For the course under consideration this leads to the report depicted in Figure 14(b). In this report we have represented the *Position in School* row using quartiles (i.e. 0-25%,25-75%,75-100%) with the top quartile rendered as +/green, and the bottom one as -/red. The *Course Change* and *Course Change Compared to School* rows are rendered with an extra device which corresponds to the standard significance levels used in hypothesis testing (1%: three stars, 5%: two stars, and 10%: one star) in both the positive and negative direction. Using this new format allows us to quickly realise that many of the unit changes depicted in Figure 14(a) were not practically significant.

Discussion

In an institutional setting, we recommend that reports similar to Figure 14(b) be used as a first reference, but with the capability to drill into the more detailed reports (i.e. Figures 12 and 13). A decision maker tasked with allocating resources to teaching teams could quickly examine a list of tables like Figure 14(b). This would also enable a prioritisation of resources e.g. the reading of free text comments as anomalous behaviour patterns are discovered, or providing support to teams that were seen to be struggling. We note that performance metrics based upon reports such as these are difficult to create, and would lose much of the rich contextual information that has been generated. However, if an institution was insistent upon following this path then we would recommend the use of reports such as these, constructed for multiple validated survey items, and displayed in a manner similar to the teacher rating forms discussed by Abrami (2001).

A number of other organisational factors could be explored in an extended model. For example, contextualising the SET responses of a cohort (i.e. in a degree) is potentially far more useful for spotting problem courses than contextualising to a school (as was done here). However, decisions such as these depend upon the underlying dataset. The Reframe dataset contains degree related information, but as one course can belong to many different degrees this is a more complex model to implement. We have chosen not to discuss this alternative in this paper for ease of communication, but such modifications are possible depending upon which questions an organisation is wishing to explore.

Beyond this, there are many ways in which the model could be extended and refined to suit different datasets or analytical questions. We leave this to other evaluation teams and to our own future work.

Conclusions

We have demonstrated that it is possible to generate a more nuanced understanding of student feedback about teaching for a single SET item. The model presented here enabled us to extract more complex information about how SET responses are distributed and change in time. This information was then condensed into a simplified format that enables quick sense-making and interpretation for decision makers and academic staff.

We acknowledge that there is a vast array of literature suggesting that using a single SET item to evaluate teaching performance is highly problematic. We agree with this literature, but institutional pressures often mandate precisely this step. We think it likely that many other evaluation units find themselves in a similar position, and so offer the techniques introduced in this paper as a way to create a more nuanced organisational dialogue.

On a wider note, it is highly surprising that the SET literature has failed to systematically adopt contemporary statistical methods, many of which have been available for decades (Gelman et al. 2013). Hypothesis testing is a fraught enterprise, and we consider it likely that many failures to demonstrate replicability of results across different organisational contexts are due an over reliance upon old fashioned techniques that were long ago abandoned by practising statisticians. In adopting a Bayesian approach we have been able to both mitigate against problems of multiplicity, and to construct a model that contains far more information than the more traditional approaches commonly reported in the SET literature. This is an important contribution, as the current era of diminishing government expenditure and increasing accountability means that more and more universities are implementing performance frameworks that make use of SET scores. In such an environment it is essential that the field investigate ways in which to reduce spurious correlations, some of which have the potential to cause considerable harm if misused.

We conclude by noting that many academic staff members are experts in the problems associated the construction of verified scales and the analysis of Likert data. This means that a poor implementation of an evaluation framework using SETs will only be met by distrust and claims of invalidity. A likely cause of the apparent failure of the SET literature to embrace more valid analytical measures lies in the silos that emerge within a university context. There is no shortage of mathematical expertise among the academy, but these experts rarely have access to the large datasets that are traditionally held by central units.

The methodology developed in this paper was made possible by a collaborative endeavour between discipline specialists from both areas (i.e. a central evaluation unit and researchers on secondment from a school of mathematics). Time was required to understand the problem, perform a thorough exploratory analysis of the data, and to construct the model. Even more time was required to develop new ways in which to enable decision makers to make sense of a modelling technique with which they are unlikely to be familiar. This was a highly unusual commitment for a university to make.

It is rare to see significant theoretical advancement when it comes to analysing institutional data. The pressures of immediately responding to short term demands leave little space for developing new ways of thinking. We hope that this work has demonstrated the value that can be obtained when a significant investment is made to encourage respectful partnerships between central units and faculty based academic staff. We encourage more institutions to make a similar commitment to investing in, supporting, and building their own collaborative relationships in the future.

Acknowledgements

The authors would like to acknowledge the feedback that has been obtained on drafts of this paper by a number of people, including Sama Low Choy, Leah Macfadyen and Shane Dawson.

References

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New directions for institutional research*, *2001*(109), 59-87.

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In *The scholarship of teaching and learning in higher education: An evidence-based perspective*(pp. 385-456). Springer, Dordrecht.

Alderman, L., & Melanie, L. (2012). REFRAME: a new approach to evaluation in higher education. *Studies in Learning, Evaluation, Innovation and Development*, *9*(1), 33-41.

Alderman, L., Towers, S., & Bannah, S. (2012). Student feedback systems in higher education: A focused literature review and environmental scan. *Quality in Higher Education*, *18*(3), 261-280.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of personnel evaluation in education*, *13*(2), 153-166.

Blaikie, N. (2003). *Analyzing quantitative data: From description to explanation*. Sage.

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

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41*, 71-88.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American psychologist*, *52*(11), 1198.

DeVellis, R. F. (2012). *Scale development: Theory and applications.* 3rd ed. Thousand Oaks, Calif: Sage publications.

Fidler, F., & Cumming, G. (2005). Teaching confidence intervals: Problems and potential solutions. *Proceedings of the 55th international statistics institute session*.

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, *53*(3), 353-374.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A. & Rubin, D. B. (1995). *Bayesian data analysis*. 3rd ed. CRC Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189-211.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American scientist*, *102*(6), 460.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337-350.

Huang, H. Y., & Wang, W. C. (2014). Multilevel higher-order item response theory models. *Educational and Psychological Measurement*, *74*(3), 495-515.

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education*, *38*(12), 1217-1218.

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education*, *5*(4), 419-434.

Kirschner, P. A., Buckingham-Shum, S. J., & Carr, C. S. (Eds.). (2012). *Visualizing argumentation: Software tools for collaborative and educational sense-making*. Springer Science & Business Media.

Krzywinski, M., & Altman, N. (2013). Points of significance: error bars.

Kynn, M. (2008). The 'heuristics and biases' bias in expert elicitation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(1), 239-264.

Low Choy, S. (2012). Priors: Silent or Active Partners of Bayesian Inference?. *Case studies in Bayesian statistical modelling and analysis*, 30-65.

Low Choy, S., & Wilson, T. (2009). How do experts think about statistics? Hints for improving undergraduate and postgraduate training.

Macfadyen, L. P., Dawson, S., Prest, S., & Gašević, D. (2016). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education*, *41*(6), 821-839.

Marsh, H. W. (1982). SEEQ: a reliable, valid, and useful instrument for collecting students'evaluations of university teaching. *British journal of educational psychology*, *52*(1), 77-95.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. *Journal of educational psychology*, *76*(5), 707.

Marsh, H. W., Hau, K. T., Chung, C. M., & Siu, T. L. (1997). Students' evaluations of university teaching: Chinese version of the Students' Evaluations of Educational Quality Instrument. *Journal of Educational Psychology*, *89*(3), 568.

Meyer, D. L. (1964). A Bayesian school superintendent. *American Educational Research Journal*, *1*(4), 219-228.

Mittal, S., Gera, R., & Batra, D. K. (2015). Evaluating the validity of student evaluation of teaching effectiveness (SET) in India. *Education+ Training*, *57*(6), 623-638.

Neumann, R. (2000). Communicating student evaluation of teaching results: rating interpretation guides (RIGs). *Assessment & Evaluation in Higher Education*, *25*(2), 121-134.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature News*, *506*(7487), 150.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... & Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

Otto, J., Sanford Jr, D. A., & Ross, D. N. (2008). Does ratemyprofessor. com really rate my professor?. *Assessment & Evaluation in Higher Education*, *33*(4), 355-368.

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: A cross-institutional comparison across 151 modules. *Computers in Human Behavior*, *60*, 333-341.

Rindermann, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education*, *42*(4), 377-399.

Rog, D. J. (2012). When background becomes foreground: Toward context-sensitive evaluation practice. *New Directions for Evaluation*, *2012*(135), 25-40.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures?. *Assessment & Evaluation in Higher Education*, *25*(4), 397-405.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*(4), 598-642.

Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education: development of an instrument based on 10 Likert-scales. *Assessment & Evaluation in Higher Education*, *32*(6), 667-679.

Sturtz, S., U. Ligges, and A. Gelman. 2005. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software* 12:1–16.

Theall, M. (2001). Can We Put Precision into Practice? Commentary and Thoughts Engendered by Abram's "Improving Judgments about Teaching Effectiveness Using Teacher Rating Forms.". *New Directions for Institutional Research*, *27*(5), 89-96.

Thompson, B. (1996). Research news and comment: AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher*, *25*(2), 26-30.

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

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

Western, B., & Jackman, S. (1994). Bayesian inference for comparative research. *American Political Science Review*, *88*(2), 412-423.

Wetzstein, M. E., Broder, J. M., & Wilson, G. (1984). Bayesian inference and student evaluations of teachers and courses. *The Journal of Economic Education*, *15*(1), 40-45.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838.

Zumrawi, A. A., Bates, S. P., & Schroeder, M. (2014). What response rates are needed to make reliable inferences from student evaluations of teaching?. *Educational Research and Evaluation*, *20*(7-8), 557-563.