Can Australian medical students' predictions of peers' responses assist with gaining reliable results on course evaluations?

S. Bacchi[1, 2], B. Guo[2, 3], S. Brown[2, 3], I. Symonds[2], J. N. Hudson[2]

Abstract

Introduction: Student feedback is integral to continuous improvement of medical programmes. A key challenge with student course evaluations is gaining large enough response rates for results to be reliable. This study investigated whether student predictions of peer, rather than personal, responses could address this challenge.

Method: An anonymous paper-based student experience of learning and teaching (SELT) survey was distributed to the Year 1–3 medical student cohorts. Students responded to 20 survey statements, using a 6-option Likert-type scale. Ten statements evaluated students' personal perspectives of the course, while the other 10 statements asked students to predict the most common response by their year cohort. Mean scores between the individual opinion-based and prediction-based statements were compared. An iterative process involving random subsampling was conducted to enable calculation of the minimum required number of responses for a stable outcome for each statement.

Results: Two hundred and fifty-nine students participated (response rate 81.7%). Three out of the 10 paired statements in the prediction-based survey accurately predicted the group opinion-based mean. For the remaining seven statement pairs, there were statistically significant (although small) differences in mean. The calculation of mean number of responses required for a stable outcome found that the prediction-based SELT required significantly fewer (189) responses than the opinion-based SELT (215) (95% CI 15.3–35.7, $p < 0.001$).

1 Queen Elizabeth Hospital, Adelaide, South Australia
2 University of Adelaide, Adelaide, South Australia
3 Royal Adelaide Hospital, Adelaide, South Australia

Correspondence
Stephen Bacchi
Medical Intern
Queen Elizabeth Hospital
28 Woodville Road
Adelaide, SA 5011
Australia
Ph: +61 8 8222 6000
Email: stephen.bacchi@sa.gov.au

ISSN 1442-1100

Conclusions: A prediction-based style of course evaluation using a 6-option Likert-type scale approximated the results gained when asking for individual opinion and required fewer responses to achieve a stable outcome.

Keywords: course evaluation; medical student; survey methods; response rate.

Introduction

In the development of a course, receiving feedback regarding the performance of the curriculum is integral to programme improvement (Kogan & Shea, 2007). Such feedback may come through both quantitative and qualitative avenues. Student evaluations, such as student experience of learning and teaching (SELT) surveys, are a commonly used way in which this feedback may be acquired (Abrahams & Friedman, 1996). One of the key challenges with student course evaluations is gaining a large enough response rate to obtain meaningful results that can be used to guide programme development. This is true for both medical and non-medical tertiary education (Fleming, Heath, Goodridge, & Curran, 2015; Guder & Malliaris, 2013). Also, frequently surveying large numbers of students may place a burden on both the students and staff (Porter, Whitcomb, & Weitzer, 2004).

There are many factors that may contribute to poor response rates to course evaluations. Such factors may include whether the evaluation is online or hard copy, evaluation length and class size (Al Kuwaiti, AlQuraan, Subbarayalu, and Piro, 2016; Crews & Curtis, 2011). Examples of strategies that may be employed in different settings to increase student response rate include grade point incentives, withholding grades until course assessment is completed, having course assessments accompany formative/summative student assessments, reminder emails and teachers/instructors using class time to promote the importance of the course evaluations (Crews & Curtis, 2011; Guder & Malliaris, 2013). Several approaches regarding increasing survey response rates in general (not specifically course evaluations) have been assessed in medical student populations, with mixed results (Grava-Gubins and Scott, 2008; Malone, Carney, House, Cranford, & Santen, 2018). The relative effectiveness of incentive-based and instructor-based strategies to improve student response rates to online course evaluations has been reviewed previously (Goodman, Anson, & Belcheir, 2014).

In an attempt to address the issues of poor response rate limiting generalisability and survey burden, Schonrock-Adema, Lubarsky, Chalk, Steinert and Cohen-Schotanus (2013) conducted a study to examine the utility of a prediction-based method of student course evaluation. This method is similar to those used in predicting election outcomes (Hofstee & Schaapman, 1990). The principle underlying this method is that each individual interacts with a group of other individuals in the same population. Therefore, the prediction of an outcome (e.g., evaluation result) by an individual not only includes their personal opinion but also represents the experience of the interactions they have had with their own subset of the population. Therefore, each individual respondent could provide a response representing that subset of the population.

ISSN 1442-1100

The prediction-based method employed in the Schonrock-Adema et al. (2013) study functions by asking students to predict the percentages of their classmates that would select each option on a Likert-type scale for a series of statements regarding the course being evaluated (rather than providing their own opinion). This study demonstrated that results representative of the cohort opinion could be gained with significantly smaller response rates with the prediction-based method (26–28 respondents with a prediction-based method compared to 67–79 respondents with a traditional method, $p < 0.0001$) (Schonrock-Adema et al., 2013). However, the use of a different modality of question/answer format, in this study, for the prediction method, in contrast with the individual opinion-based method, has been challenged (Dolmans, Kamp, Stalmeijer, Whittingham, & Wolfhagen, 2014).

This project aimed to explore whether, using the same answer modality for the two evaluation methods (in this instance a 6-option Likert-type scale to respond to all survey statements), Australian student predictions of course evaluations could (a) accurately predict cohort opinion and (b) achieve a stable outcome with fewer respondents than necessary when asking for individual opinion.

Method

Participants

Students from years 1–3 of the University of Adelaide, Bachelor of Medicine, Bachelor of Surgery programme (6-year school-entry undergraduate degree) participated in the evaluations. Only students who attended the recruitment lectures, at which the surveys were distributed by student investigators and independent administrators, were able to participate. Participation was voluntary.

Surveys

Students were invited to complete an anonymous paper-based survey prior to scheduled medical programme lectures. Surveys were distributed near the end of the academic year (October 2017). The survey included demographic questions and 20 statements requiring a response regarding the programme's tutorials, using a 6-option Likert-type scale. These 20 statements were divided into two sets of 10, with each set containing the same statements (see Table 1). However, one set of 10 was preceded with the question "To what extent do *you* agree/disagree with the following statements?" (opinion-based response) and the other was preceded by "Do you think that *the majority of your cohort would respond that they* agree/disagree with the following statements?" (prediction-based response).

ISSN 1442-1100

Table 1
Questions Included in the Opinion-Based and Prediction-Based SELTs, With Statistical Significance of Difference Between Opinion-Based and Prediction-Based Mean Scores

	Statistical significance of comparison between opinion-based SELT[1] mean and prediction-based SELT[2] mean scores	**Confidence interval of difference between mean scores**
My general impression of the tutorials is positive.	< 0.001*	0.18–0.48
The tutorial cases have an appropriate number of sessions.	0.244	-0.05–0.22
The tutorial sessions are an appropriate length (i.e., duration of each tutorial).	< 0.001*	0.17–0.42
The frequency of tutorial sessions is appropriate (i.e., number of tutorials per week).	0.002*	0.08–0.36
The "prompts for students" on handouts are NOT relevant to the content of the cases.	< 0.001*	0.28–0.6
The tutorial cases follow a logical sequence.	0.001*	0.10–0.38
The tutorial cases have clear learning objectives.	0.096	-0.03–0.38
The lectures were NOT relevant to my learning for tutorial cases.	0.952	-0.18–0.17
The online lecture content was relevant to my learning for tutorial cases.	< 0.001*	0.13–0.42
The provided tutorial formative self-assessment questions are useful.	0.004*	0.08–0.41

1 The opinion-based SELT questions were preceded by the question: To what extent do you agree/disagree with the following statements?

2 The prediction-based SELT questions were preceded by the question: Do you think that the majority of your cohort would respond that they agree/disagree with the following statements?

* $p < 0.05$.

ISSN 1442-1100

Analysis

Values from one to six were allocated to the Likert-type responses. The two negatively-worded questions in each of the surveys (see Table 1) were reverse-scored (with 6 points allocated to "very strongly disagree" and 1 point to "very strongly agree"). The means of each of the 10 paired statements were compared to determine whether the students had accurately predicted the cohort's responses. Results were analysed using RStudio (version 1.0.153). Parametric statistics, namely unpaired t-tests, were used for the comparison of the Likert-type responses. The use of parametric statistics with Likert-type questions has been debated previously (Carifio & Perla, 2008; Sullivan & Artino, 2013; Wadgave & Khairnar, 2016). An approach similar to that employed in the Schonrock-Adema et al. (2013) study was used to determine whether the prediction-based style of question enabled a stable outcome with a smaller number of respondents compared to opinion-based responses. For this approach, the following iterative process was performed independently for each statement in the opinion-based responses and prediction-based responses.

For an individual statement, the iterative process began by selecting a random response (subsample). The percentage of responses per Likert-option was then calculated. These percentages (subsample percentages per Likert-option) were then compared to those percentages per Likert-option obtained with the total sample (total sample percentages per Likert-option). The difference between each subsample option percentage and total sample percentage was then calculated (e.g., if 50% of respondents responded "strongly agree" in the subsample and the total sample option percentage responding "strongly agree" was 30%, then the percentage difference for that option would be 20%).

All percentage differences were then added together (e.g., percentage difference for very strongly agree through to very strongly disagree) to calculate the aggregate percentage difference. If the aggregate percentage difference was > 5%, then another random response was selected for that statement and the above calculations repeated.

The process was repeated until the aggregate percentage difference was consistently < 5%. The number of responses in the subsample at this point provided the number of responses required for a stable outcome.

The iterative process was then repeated five times for each statement. The total mean number of responses required for a stable outcome for both opinion-based (personal) and prediction-based responses were then compared using unpaired t-tests.

Pooled student responses (both opinion-based and prediction-based) were made available to faculty on completion of the survey to inform subsequent course development. No individual responses were made available to members of faculty. The University of Adelaide Ethics Committee granted approval for this project (H-2017-092).

ISSN 1442-1100

Results

In total, there were 259 participants (81.7% of the 317 students who attended recruitment lectures). There were 84 first-year participants (83.2%), 96 second-year (87.3%) and 79 third-year (74.5%). There were 148 female participants (57.1%).

Accuracy of student predictions

When comparing the mean agreement scores to the opinion-based SELT statement and prediction-based equivalents, there were three statements in which there was no statistically significant difference between the opinion-based agreement mean score and the prediction-based agreement mean score (see Table 1). These results indicate that for these three statements, the students accurately predicted the group response.

For the other seven statements, there was a statistically-significant difference between the mean response to the opinion-based SELT and the prediction-based SELT. While several of these comparisons found highly statistically-significant results ($p < 0.001$), the magnitude of the difference between the opinion-based SELT response and prediction-based SELT response was typically quite low. For example, the statement in which there was the greatest discrepancy between opinion-based and prediction-based responses was "My general impression of the tutorials is positive". The statistical significance of the difference between the means of the opinion-based and prediction-based responses was $p < 0.001$. However, the confidence interval for the difference between the means was 0.18–0.48. All other confidence intervals for the mean difference on other statements were closer to zero than for this statement (see Table 1). Accordingly, it can be seen that although statistically significant differences exist, the small effect size of these differences may mean that these differences do not bear major significance on the practical interpretation of the survey results.

Number of responses required for stable outcome

Using the method outlined in the method section, the mean number of responses required for a stable outcome was calculated for each of the opinion-based and prediction-based statements. The mean number of responses required for a stable outcome in the prediction-based SELTs (189 responses, or a response rate of 73%) was significantly lower than that for the opinion-based SELTs (215 responses, or a response rate of 83%). These results indicate that for the given set of statements, on average, 26 fewer responses were required for a stable outcome with the prediction-based SELT compared to the opinion-based SELT (95% CI 15.3–35.7, $p < 0.001$).

Discussion

This study has found that prediction-based SELTs using a 6-option Likert-type scale may reasonably accurately estimate the total mean individual opinion-based SELT response to evaluate a series of statements (see Table 1). Calculation of the mean number of responses required for a stable outcome revealed that the prediction-based SELT required a response rate of 73%. This required response rate was significantly lower than that required for a stable outcome using the individual opinion-based SELT.

ISSN 1442-1100

The findings of this study are consistent with studies that have used predictive methods of course evaluation in medicine previously, with all results indicating that the use of prediction-based methods of evaluation can achieve stable outcomes with fewer respondents than individual opinion-based methods (Cohen-Schotanus, Schonrock-Adema, & Schmidt, 2010; Schonrock-Adema et al., 2013). The previous studies in this area have used 4-option Likert-type scales, whereas the current study employed a 6-option Likert-type scale. The results of this study indicate that the prediction-based course evaluation method may be generalisable to other styles of Likert-type questions.

It was interesting that in the statements that had a statistically significant difference between the prediction-based and opinion-based results, the student predictions always underestimated the positivity of the group's opinion. In other words, the actual group opinion was more positive than students predicted it would be. This was the case even in the negatively-worded statements for which students predicted that the group would agree more strongly than they in fact did. Given this pattern, it is possible that a standard correction could be employed on the results of prediction-based SELTs that may make their answers a more accurate approximation of the individual opinion-based SELTs.

A limitation of this project was that it was conducted at a single centre, only in English. Given the importance of semantics in the wording of the questions in this study, further studies in other languages would be required prior to generalising the findings to centres at which languages other than English are spoken. It is important to note that this project involved only hard-copy surveys. Many course evaluations are now conducted online. It is possible that conducting such surveys in an online format would influence the outcomes of the project. All participants within each cohort completed both prediction-based and opinion-based surveys rather than being randomly assigned a survey type, in order to gain a larger sample size (i.e., so that the whole cohort could complete both types of survey rather than only half the cohort). The pattern of results appeared consistent across the three year levels; however, statistical analyses were not conducted between year levels. Student investigators were involved in the distribution of surveys due to logistical challenges. This is unlikely to have introduced bias since both the individual opinion-based SELTs and the prediction-based SELTs were distributed by the same personnel. Given that multiple university courses conduct programme and course evaluation via SELTs, it should be noted that this study was conducted solely with medical students.

It should also be noted that while this study does not have the issue of differing question modalities raised by Dolmans et al. (2014), no single modality should be used for course evaluation. Ideally qualitative and quantitative methods should be used in combination. The comment by Parker (2013) regarding the importance of continuing to investigate models outside of the Kirkpatrick model for programme evaluation is also worth noting.

Further studies of prediction-based evaluation methods are required prior to their implementation. These studies may benefit from larger sample sizes and from being conducted across multiple sites. Future research may also seek to investigate whether a standard correction for the group predictions may improve the accuracy of individual

ISSN 1442-1100

opinion estimates. Mixed methods involving focus groups conducted after prediction-based surveys may also be useful in examining the reasons underlying any differences in opinion-based and prediction-based answers. It would also be useful to examine if a prediction-based method of course evaluation may be effective in non-medical or non-health courses.

Conclusions

Prediction-based SELTs using a 6-option Likert-type scale may reasonably accurately estimate the total mean response on individual opinion-based SELTs. Such an approach may produce a stable outcome from a smaller number of respondents than when assessing individual opinion. Further studies of such prediction-based evaluation methods are required prior to their implementation.

References

Abrahams, M., & Friedman, C. (1996). Preclinical course-evaluation methods at U.S. and Canadian medical schools. *Academic Medicine*, *71*(4), 371–374.

Al Kuwaiti, A., AlQuraan, M., Subbarayalu, A. V., & Piro, J. S. (2016). Understanding the effect of response rate and class size interaction on students evaluation of teaching in a higher education. *Cogent Education*, *3*(1). doi:10.1080/2331186x.2016.1204082

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, *42*(12), 1150–1152. doi:10.1111/j.1365-2923.2008.03172.x

Cohen-Schotanus, J., Schonrock-Adema, J., & Schmidt, H. G. (2010). Quality of courses evaluated by "predictions" rather than opinions: Fewer respondents needed for similar results. *Medical Teacher*, *32*(10), 851–856. doi:10.3109/01421591003697465

Crews, T. B., & Curtis, D. F. (2011). Online course evaluations: Faculty perspective and strategies for improved response rates. *Assessment & Evaluation in Higher Education*, *36*(7), 865–878. doi:10.1080/02602938.2010.493970

Dolmans, D., Kamp, R., Stalmeijer, R., Whittingham, J., & Wolfhagen, I. (2014). Biases in course evaluations: "What does the evidence say?". *Medical Education*, *48*(2), 219–220. doi:10.1111/medu.12297

Fleming, P., Heath, O., Goodridge, A., & Curran, V. (2015). Making medical student course evaluations meaningful: Implementation of an intensive course review protocol. *BMC Medical Education*, *15*, 99. doi:10.1186/s12909-015-0387-1

Goodman, J., Anson, R., & Belcheir, M. (2014). The effect of incentives and other instructor-driven strategies to increase online student evaluation response rates. *Assessment & Evaluation in Higher Education*, *40*(7), 958–970. doi:10.1080/02602938.2014.960364

ISSN 1442-1100

Grava-Gubins, I., & Scott, S. (2008). Effects of various methodologic strategies: Survey response rates among Canadian physicians and physicians-in-training. *Canadian Family Physician*, *54*(10), 1424–1430.

Guder, F., & Malliaris, M. (2013). Online course evaluations response rates. *American Journal of Business Education*, *6*(3), 333–338.

Hofstee, W., & Schaapman, H. (1990). Bets beat polls: Averaged predictions of election outcomes. *Acta Politica*, *25*, 257–270.

Kogan, J. R., & Shea, J. A. (2007). Course evaluation in medical education. *Teaching and Teacher Education*, *23*(3), 251–264. doi:10.1016/j.tate.2006.12.020

Malone, M. G., Carney, M. M., House, J. B., Cranford, J. A., & Santen, S. A. (2018). Tit-for-tat strategy for increasing medical student evaluation response rates. *Western Journal of Emergency Medicine*, *19*(1), 75–79. doi:10.5811/westjem.2017.9.35320

Parker, K. (2013). A better hammer in a better toolbox: Considerations for the future of programme evaluation. *Medical Education*, *47*(5), 440–442. doi:10.1111/medu.12185

Porter, S., Whitcomb, M., & Weitzer, W. (2004). Multiple surveys of students and survey fatigue. *New Direction for Institutional Research*, *121*, 63–73.

Schonrock-Adema, J., Lubarsky, S., Chalk, C., Steinert, Y., & Cohen-Schotanus, J. (2013). "What would my classmates say?" An international study of the prediction-based method of course evaluation. *Medical Education*, *47*(5), 453–462. doi:10.1111/medu.12126

Sullivan, G., & Artino, A. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. doi:0.4300/JGME-5-4-18

Wadgave, U., & Khairnar, M. R. (2016). Parametric tests for Likert scale: For and against. *Asian Journal of Psychiatry*, *24*, 67–68. doi:10.1016/j.ajp.2016.08.016

ISSN 1442-1100