2009

Response style contamination of student evaluation data

Sara Dolnicar
University of Wollongong, s.dolnicar@uq.edu.au

Bettina Grun
Wirtschaftsuniversitat Wien, bettina@uow.edu.au

Publication Details

Dolnicar, S. & Grun, B. (2009). Response style contamination of student evaluation data. Journal of Marketing Education, 31 (2), 160-172.

Response style contamination of student evaluation data

Abstract
Student evaluation surveys provide instructors with feedback regarding development opportunities and they form the basis of promotion and tenure decisions. Student evaluations have been extensively studied, but one dimension hitherto neglected is the actual measurement aspect: which questions to ask, how to ask them, and what answer options to offer to students to get the most valid results. This study investigates whether cross-cultural response styles affect the validity of student evaluations. If they do, then the student mix in a class can affect an instructor's evaluation, potentially producing biased feedback and prompting inappropriate decisions by university committees. This article discusses two main response styles, demonstrates the nature of the bias they can cause in student evaluation surveys using simulated artificial data, and illustrates three cases based on real student evaluation data in which marketing instructors' teaching quality assessments may be heavily biased because of response styles. The authors propose a simple method to check for response style contamination in student evaluation data and they discuss some practical implications.

Keywords
data, style, contamination, student, evaluation, response

Disciplines
Business | Social and Behavioral Sciences

Publication Details
Dolnicar, S. & Grun, B. (2009). Response style contamination of student evaluation data. Journal of Marketing Education, 31 (2), 160-172.

Response style contamination of student evaluation data

Abstract

Student evaluation surveys provide instructors with feedback regarding development opportunities, and form the basis of promotion and tenure decisions. Student evaluations have been extensively studied, but one dimension hitherto neglected is the actual measurement aspect: which questions to ask, how to ask them and what answer options to offer to students to get the most valid results. This study investigates whether cross-cultural response styles affect the validity of student evaluations. If they do, then the student mix in a class can affect an instructor's evaluation, potentially producing biased feedback and prompting inappropriate decisions by university committees.

This paper discusses two main response styles, demonstrates the nature of the bias they can cause in student evaluation surveys using simulated artificial data and illustrates three cases based on real student evaluation data in which marketing instructors' teaching quality assessments may be heavily biased because of response styles. We propose a simple method to check for response style contamination in student evaluation data, and discuss some practical implications.

Keywords: student evaluations, survey, answer format, Likert scale, ordinal, response style, satisfaction

INTRODUCTION

Student evaluations have been an integral part of teaching at universities for many decades. According to the Carnegie Foundation (quoted in Babin, Shaffer & Tomas, 2002), 98 per cent of universities use systematic student evaluations. Surveys give teachers valuable feedback, which helps them improve their teaching and the quality of the service provided to the students. Student evaluations also affect academics' careers, tenure and promotion prospects. Simpson and Siguaw (2000, p. 199) argue that: "Student evaluation of teaching instruments are commonly administered by universities to presumably provide feedback to faculty for improvement of teaching effectiveness. Instead, these measures are routinely used as a basis for determining faculty merit, promotion and tenure, making the instrument vitally important to faculty."

Unfortunately, student evaluations are not designed for comparative use, which is typically what promotion and tenure committees want to use them for. Wilson warned of this as far back as 1982: "The area of administrative use of evaluations is the area most urgently requiring validation because decisions affect lives and careers" (Wilson, 1982, p. 9). Wilson called for more measurement-related research, especially in the field of marketing where questionnaires are a typical measurement tool, the quality of which determines the validity of research findings. This paper responds to the call for further measurement-related research into student evaluations. More specifically, it investigates whether and how cross-cultural response styles can lead to biased conclusions in the assessment of student evaluations. We provide empirical illustrations of conclusions biased due to response styles based on real student evaluation data, and propose a simple method for checking for response style bias in student evaluation data sets to avoid invalid conclusions.

The paper is structured as follows: we first review prior research in the area of methodology and measurement-related student evaluation research, as well as response style research. Next, we introduce a simple procedure that can be used to assess the possible bias in any given student evaluation data set. We then use this procedure to demonstrate — using a simulated artificial data set — the kind of bias that can result from cross-cultural response styles, and provide three empirical examples of how a marketing lecturer's teaching quality assessment can be biased by cross-cultural response styles. These examples are based on real student evaluation data. Finally, we draw practical conclusions for marketing education.

PRIOR RESEARCH — STUDENT EVALUATIONS

Prior research can be classified into three categories: (1) studies assessing particular evaluation tools, (2) studies identifying factors that affect student evaluation results and (3) method and measurement-related studies. Because our study contributes to the last of the three areas, we also limit our literature review to method and measurement-related studies.

Only a minority of studies that investigate student evaluations focuses on methodological and measurement aspects, and within these, the following problems have been identified: Student evaluations do not differentiate well between faculties (Wheeler and Geurts, 1986). They only distinguish between the very best and the very worst teachers (Wheeler and Geurts, 1986), and they capture "halo effects", which decrease as the overall rating of the instructor increases (Orsini, 1988). The validity of student evaluations is reduced by several language difficulties: (1) people do not have the necessary level of awareness to evaluate all aspects of a situation, (2) disagreement may exist regarding the object of the evaluation, (3) denotative and connotative

meanings cannot be separated, (4) signs and meanings may be inconsistent and (5) the situation may affect the evaluation (Bertsch and Peek, 1982). Grunenwald and Ackerman (1986) and Spivey and Caldwell (1982) express similar concerns about the meaning of items in evaluations. Prior research also identifies significant disparities in student satisfaction — even amongst those who work in the same group and receive the same final mark — which questions the validity of student evaluation surveys (Appleton-Knapp and Krentler, 2006). Furthermore, the instructor's personality may explain most of the teaching evaluation score — twice as much as the second-strongest factor (Clayson and Haley, 1990; Clayson and Sheffret, 2006). Expected grade, the instructor's personality and the instructor's likeability can account for 73 per cent of the explanation of the evaluation (Clayson and Sheffret, 2006). Prior research also identifies that the instructor's experience is not captured by repeated student evaluations (Clayson, 1999), and assuming that instructors do use evaluations to improve their teaching, this undermines the validity of student evaluation measures.

Despite the wide range of methodological and measurement concerns raised by researchers in the past, which all indicate that student evaluation results cannot be taken at face value, no study has investigated the effects of cross-cultural response styles on student evaluation results, even though response styles have been extensively studied in marketing research, and some studies have investigated the validity of student evaluation instruments in different countries and cultural backgrounds (for example, Marsh et al., 1997; 1998). These studies compared the psychometric properties of the instrument, rather than comparing the responses directly; response style bias is unlikely to occur in such a situation. Furthermore, evaluations of the best and the worst teachers were used as the basis of analysis. Cross-cultural response styles may not affect such extreme

judgements as dramatically as they affect evaluations of teachers who are not perceived as being at the extremes of the teacher quality continuum.

PRIOR RESEARCH — RESPONSE STYLES

Response styles — specific kinds of response biases — are the key object of investigation in this study. A response bias is "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (that is, what the items were designed to measure)" (Paulhus, 1991, p. 17). Response styles are response biases that "an individual displays… consistently across time and situations" (Paulhus, 1991, p. 17). This paper uses the term *response style* with this definition throughout.

A range of different response styles exists (Baumgartner and Steenkamp, 2001), but this study focuses on the two major forms: extreme response style (ERS) and acquiescence response style (ARS). ERS occurs when respondents prefer to use the endpoints of the available answer options (such as "strongly agree" and "strongly disagree" on a multi-category answer format such as the frequently used Likert scale). ARS occurs when respondents avoid giving negative answers and tend to use positive answer options (such as "strongly agree", "agree" and "slightly agree").

One factor consistently identified as being associated with the occurrence of response styles is the cultural background of respondents (Chun, Campbell & Yoo, 1974; Hui & Triandis, 1989; Marin, Gamba & Marin; 1992; Marshall & Lee, 1998; van Herk, Poortinga & Verhallen; 2004; Welkenhuysen-Gybels, Billiet & Cambre, 2003; Zax & Takahashi, 1967). Findings indicate that

Americans, Australians and Hispanic respondents are characterised by ERS, whereas Asian respondents tend to have a "mild" response style, avoiding the endpoints — the opposite of ERS.

If ignored, the presence of response styles will affect the validity of empirical research findings (Dolnicar & Grün, 2007b). ERS spuriously increases reliability, decreases validity (Clarke III, 2001) and produces a more extreme frequency distribution. Consequently, standard deviations increase and correlations decrease, and this affects all methods that use correlation analysis as their foundation, such as factor analysis or regression analysis (Chun, Cambell & Yoo, 1974; Hui & Triandis, 1989; Heide & Gronhaug, 1992) — two of the most commonly used techniques in empirical social sciences research.

ARS can lead to a separate factor in factor analyses, which essentially represents an artefact because it contains only negatively-keyed variables (Heide & Gronhaug, 1992). The presence of respondents tending to prefer either upper or lower answer options leads to spuriously higher correlations; thus, covariance-based analyses can be substantially influenced (Rossi, Gilula & Allenby, 2001). Analytic techniques based on distance computations are also affected, for example, cluster analysis, which is frequently used to identify homogeneous sub-groups. Because similarities and dissimilarities between responses form the basis of clustering, results can be heavily biased (Greenleaf, 1992a).

Traditionally, two fundamentally different ways of addressing the problem of response styles exist. The better of the two options is to collect data in a way that minimises the risk of response style contamination. This can be achieved, for example, by using binary (yes/no) answer formats (Cronbach, 1950), or by using best-to-worst scaling (Lee, Soutar, Louviere & Daly, 2006). If the required questions cannot be asked using either of the above answer formats, or

because the data set has already been collected, the second option must be implemented: correction for response styles. Although several researchers recommend this approach (Byrne & Campbell, 1999; Cheung & Rensvold, 2000; Fischer, 2004; Greenleaf, 1992a; 1992b; Van de Vijver & Poortinga, 2002; Welkenhuysen-Gybels, Billiet & Cambre, 2003), a major danger is associated with the procedure: no guarantee exists that the correction method chosen is correct and will eliminate response styles contamination. At worst, new, undesired contamination could even be introduced.

To assess whether contamination of survey data with response styles is merely a theoretical problem, or whether it affects a significant proportion of empirical survey data sets used in satisfaction research in general and student evaluation surveys in particular, we conducted two comprehensive literature reviews: one including all relevant articles (31) published in *Managing Service Quality* between 2000 and 2007 and one including all relevant articles (9) published in the *Journal of Marketing Education*.

Results from the review of studies published in *Managing Service Quality* show that only 13 per cent of service satisfaction studies were based on multicultural data sets, while 68 per cent used multi-category ordinal response options to measure satisfaction. These findings prompted us to conclude that service satisfaction studies in general are not likely to suffer from lack of validity due to cross-cultural response styles. However, the heavy use of multi-category answer options could introduce individual level response style bias. No study examined acknowledges the possibility of such data contamination.

Student evaluation studies published in the *Journal of Marketing Education* rarely provide information about student composition with respect to culture. The only study to offer details

about the origin of students was a cross-county evaluation survey by Clarke III and Flaherty (2002) containing 177 responses from students from China, the UK and the US. The authors account for cultural difference in meaning, but do not mention possible dangers of data contamination by culture-specific response styles. Regarding other factors affecting response style contamination, the studies exhibit significant variability in sample size, number of subjects, number of instructors and evaluation tools used; yet all studies (without exception) use ordinal answer formats. These results indicate that student evaluation surveys are at risk of stimulating the use of response styles in respondents, especially if the respondents come from different cultural backgrounds.

A PROCEDURE FOR IDENTIFYING POTENTIAL RESPONSE-STYLE BIAS

The exact nature of the bias caused by response styles is not known. We propose a procedure that considers several possible alternative sources of response-style bias. We take the originally collected (raw) data set as a starting point, and compute several "corrected" data sets. Each "correction" assumes the presence of a different kind of bias (ERS only, ARS only, both ARS and ERS). We then conduct data analysis (for example, testing whether the evaluations of two instructors differ) on all those data sets. If the results are the same for all data sets, we can assume that response styles have not biased the data to a significant extent. However, if different correction methods lead to different conclusions, we can assume that the data is biased, and comparative evaluation results for the raw data may not be valid.

The details of each step of the procedure are explained below.

Step 1: Is there empirical evidence of cross-cultural response styles in the data?

We first assess whether any empirical evidence of cross-cultural response styles in the data exists. If none is found, data analysis can safely be conducted without following the remaining steps of the proposed procedure.

A simple way of answering this question is to use prior knowledge about occurring response styles. For instance, if previous work shows that Asian respondents tend to use the middle answer categories ("mild" response style), while US respondents tend to use the extreme answer options (ERS), and if the student sample contains responses from both Asian and US students, simple exploratory analysis can be conducted to test whether these two groups show the assumed response styles. We can compute frequency counts of Asian and US student responses for each answer category and use a chi-squared test to assess whether differences in frequency distribution are random or statistically significant (see Table 2 for an example of such an analysis). If statistically significant, response styles are present, and further investigations of the data are required.

Step 2: Which methods are suitable to correct for the response styles in the data?

Many methods have been proposed in the past to correct for response styles. Not all are suitable for each response style. Consequently, we should select a subset of suitable correction techniques (for guidance on suitable correction methods see for standardisation techniques for example Fischer, 2004; and for model-based approaches see for example Rossi, Gilula & Allenby, 2001, Johnson, 2003 and De Jong et al., 2008). The results from Step 1 indicate which

response styles are present, for example, ARS. In the case of ARS, the positive bias of respondents with a stronger ARS than others can be corrected by subtracting the individual mean estimate over all answers a respondent has given.

Once a subset of suitable techniques is identified, each technique is applied to the raw data set separately, leading to several *derived data sets*, as illustrated in Figure 1 by the grey boxes.

---------- please insert Figure 1 approximately here ----------

All subsequent steps use multiple, different correction methods and the original data set, because the true nature of contamination is unknown. Any single correction method might either fail to correct for the response style present — or even lead to the introduction of further bias into the data.

Step 3: Do all data sets lead to the same conclusion about student evaluations?

Step 3 is the actual data analysis step. For example, the university may choose to compute an average across all student evaluation survey questions for one instructor and compare it with the overall university average. When response styles are present in the data, this computation is run for the raw data, but also for all derived data sets. If, for example, three derived data sets emerged from Step 2, four separate and independent computations are conducted: one with the

raw data and one with each of the three derived data sets, as indicated by the white box in Figure 1.

When the results for all data sets are available, they can be compared. In the best possible case, the same conclusions will be drawn. For example, that instructor X is found to achieve significantly better results than the average instructor at that university in all four cases. Such stable results are considered valid and reliable and can safely be used as a basis for determining faculty merit, promotion and tenure, as well as for improving teaching effectiveness. All other conclusions must be treated with care, and may require additional investigation before using them as the basis for decisions on promotion or tenure or for guidance regarding teaching improvement. For instance, two of the four computations may indicate that instructor X is better than the average instructor at the university, while two may lead to the opposite conclusion. This last stage of the procedure is illustrated by the black boxes in Figure 1.

This procedure has been used successfully in tourism research, an area particularly prone to cross-cultural response style contamination, because most respondents are tourists from different cultural backgrounds (Dolnicar & Grün, 2007a; Dolnicar, Grün & Le, in press).

ILLUSTRATION WITH ARTIFICIAL DATA

The main problem with response styles and research into response styles is that the researcher never knows whether response styles are present, which kinds of response styles may be present and how they affect the results. To provide solid evidence of the mechanisms discussed above, one of two kinds of data is needed: experimental data with students or simulated

artificial data. Because it is impossible to collect real student evaluations following a strict experimental design that allows manipulation of all variables assumed to affect the evaluation (for example, the same instructor assessed under the exactly same conditions except for the cultural mix of students twice), we constructed an artificial data set modelled on real student evaluation data to demonstrate the effects of cross-cultural response styles. Heide and Gronhaug (1992) also used a simulation study with artificial data to show the impact of response styles on data analysis results.

Artificial data is constructed to demonstrate how response styles affect the comparison of course evaluations. The generation of artificial data is based on procedures previously proposed, where the answers to questions are assumed to be represented in the respondents' minds by latent continuous variables which they must map to the ordinal, multi-category scale when answering the questionnaires. Differences in response styles occur due to different mapping functions which respondents use. The Appendix contains technical details of how these data characteristics were modelled in the simulated artificial data.

The characteristics of the simulated artificial data set are as follows: Two instructors (two classes) were each evaluated by 50 students. The students stated their level of agreement with seven questions relating to perceived teaching quality on a six-point multi-category scale for each course. The students perceived the instructors as equally good on the first three items, the first instructor as better on the next two attributes and the second instructor as better on the last two attributes. Consequently, the correct conclusion form the student evaluation analysis is that both instructors were equally good on the first three items, instructor 1 outperformed instructor 2 on the next two questions and instructor 2 outperformed instructor 1 on the last two questions. We

refer to this outcome as the "correct assessment of teaching quality" for the analysis using artificial data.

While the perceptions of the students were the same within each course (that is, the latent continuous variables have the same mean value for each question and course), one subgroup of students (local students) demonstrated an ERS, one subgroup (international students) demonstrated an ARS in responding to the survey questions, that is, these subgroups differed with respect to the mapping of the latent continuous variable onto the ordinal, multi-category scale. The composition of classes with respect to these two subgroups is the factor varied in the simulation study. We compared the evaluation of the teaching quality of the two instructors under four different conditions, where: (1) both classes consisted only of local students, (2) both classes consisted only of international students, (3) both classes consisted of 50 per cent local students and 50 per cent international students and (4) the first class consisted of local students only and the second class consisted of international students only. Data was generated independently 100 times according to this sampling scheme and the procedure was applied to each of the 100 data sets. Results were benchmarked against the "correct assessment of teaching quality" as defined in the previous paragraph. For each scenario, the number of correct results could be between 0 (zero) and 700 (100 computations with seven questions each). Higher values indicated a higher proportion of correct results.

We used standardisation methods to correct for response bias. Standardisation is the most commonly used technique to adjust for response styles in cross-cultural research (Fischer, 2004). Because both ARS and ERS were present in the artificial data, derived data sets were constructed using the following two correction methods: (1) substraction of individual means (to account for ARS) and (2) substraction of individual means and division by the individual standard deviations

(to account for both ERS and ARS). These corrections were made on an individual and a group-wise level (implied by the international student indicator) using all available evaluations for each student.

Table 1 contains the results for the four comparisons. Each row provides the results for one scenario. Each column represents one data set (either raw data or one of four different correction methods accounting for different response styles). The numbers indicate the percentage of the 700 comparisons which led to agreement with the "correct assessment of teaching quality".

---------- please insert Table 1 approximately here ----------

As evident from the table, all data sets lead to high agreement with the "correct assessment of teaching quality" for three of the four scenarios, namely those scenarios where response styles do not affect the results because either only students from one cultural background are included, or the proportion of international and local students is the same in both classes. In the fourth scenario, which is characterised by students from different cultural backgrounds (with different response styles) attending each of the two classes, the number of "correct assessments of teaching quality" drops dramatically. If only the raw data is used, a correct assessment is made in only 71 per cent of cases; if the data is used that corrects only for ARS, 84–85 per cent of assessments are correct; if both ARS and ERS are accounted for, 97 per cent of cases are assessed correctly. We know the nature of the bias for these data sets; where the nature of bias is not known (for example, real student evaluation data), the proposed analysis would still reveal that three of four

comparisons can be made safely, but the comparison of instructor 1 and instructor 2 in the fourth scenario (classes with a different student mix regarding cultural background) is biased.

EMPIRICAL ILLUSTRATION

The previous section demonstrated the bias that can result from different class compositions with respect to cultural backgrounds of students. We now illustrate — using real student evaluation data — three situations in which the assessment of a marketing instructor's teaching performance can be biased because cross-cultural response styles are not taken into consideration: (1) the comparison of a marketing instructor teaching a postgraduate subject with an instructor from another faculty teaching a postgraduate subject, (2) the comparison of a marketing instructor teaching a postgraduate subject with an instructor from the same faculty, but another discipline (accounting) teaching a postgraduate subject and (3) the comparison of a marketing instructor teaching a postgraduate subject with another marketing instructor teaching an undergraduate subject.

Data

The student evaluation data used in this illustration was collected at an Australian university. Australia generally has a high proportion of international students; consequently, we would expect cross-cultural response styles to affect the validity of student satisfaction measurement.

At the University of Wollongong, where the data for this illustration was collected, student evaluations are conducted upon request by the instructor by an independent survey administrator.

The survey administrator invites the students to complete the evaluation using a standard set of written instructions. The instructor is not present while the evaluation takes place, and the envelope containing the evaluations is collected at the end of the exercise and submitted by the independent survey administrator. This process ensures that the instructor cannot influence the results or any possible systematic ways of answering the questions (such as response styles).

The data used for the empirical investigation was collected in March, April and May of 2006, and consisted of 6,844 fully completed questionnaires completed in 883 different subjects by 2,489 different students across all disciplines. On average, each student completed the questionnaire 2.7 times for different subjects.

Survey instrument

The student evaluation questionnaire contained seven questions. The precise wording was: "The learning objectives for this subject were made clear", "The criteria for assessment in this subject were made clear", "I have developed a good understanding of the content of this subject", "My learning in this subject was well supported", "This subject helped me to think critically/analytically", "As a result of my experience with this subject I am enthusiastic about further learning" and "Overall, I am satisfied with my learning experience in this subject".

Students used a typical six-point Likert-type scale which was fully labelled as "strongly agree", "agree", "slightly agree", "slightly disagree", "disagree" and "strongly disagree". Equidistant numerical scores, from -1 for "strongly disagree" to 1 for "strongly agree" formed the raw data set. Additional information for each student was available on their international

indicator status at university, their citizenship and their language spoken at home. This information was not contained in the survey itself. It was included *ex post*, using student record information which was linked to the survey data sets using student ID numbers.

Sample characteristics

Of the 6,844 completed student evaluation questionnaires, 5,481 (80 per cent) were completed by 2,024 local students, and 1,363 (20 per cent) by 465 international students. Among local students, 98 per cent were Australian citizens. International students came from a wide range of countries of origin: 49 per cent were Chinese, 10 per cent were from the USA, four per cent from Hong Kong and more than two per cent of students were Canadian, Taiwanese, Indonesian, Thai, Indian and Singaporean. The distribution of students from different cultural backgrounds was extremely skewed: 581 subjects (66 per cent) contained less than one-fifth international students, while 133 subjects (15 per cent) had at least 80 per cent international students.

Assessing response style contamination

Step 1: Is there empirical evidence of cross-cultural response styles in the data?

The average use of each of the six answer options ("strongly disagree" to "strongly agree") was computed separately for (1) the local students, (2) all international students and (3) Chinese

students (see Table 2). Differences in response styles became visible if groups of respondents demonstrate significantly different frequencies of using each one of the available answer options.

---------- please insert Table 2 approximately here ----------

The association between international student status and use of answer options ($\chi^2 = 149.4$, df = 5, p-value < 0.01) is highly significant, indicating that international students tend to avoid extreme answer options and agree with satisfaction statements. The association between local students and Chinese students is also significant ($\chi^2 = 129.2$, df = 5, p-value < 0.01).

A second way to assess the existence of response styles is by analysing individual means and standard deviations over a range of questions. For each student, the answers for all subjects were combined to determine these values. Results indicate that the international students had significantly higher mean values (Welch *t*-test: t = -2.7, df = 740.4, p-value < 0.01) and smaller standard deviations (Welch *t*-test: t = 8.1, df = 764.5, p-value < 0.01) than the local students. Chinese students also differed significantly from local students (Welch *t*-test for means: t = -2.8, df = 285.1, p-value < 0.01; Welch *t*-test for standard deviations: t = 8.7, df = 294.5, p-value < 0.01). A comparison of all international students with Chinese students indicates that Chinese respondents had smaller individual standard deviations (Welch *t*-test: t = 2.3, df = 447.7, p-value < 0.03).

Based on the analyses of frequencies, means and standard deviations, the key question of Step 1 (Is there empirical evidence of cross-cultural response styles in the data?) must be answered "yes" — the student evaluation data set does contains cross-cultural response styles.

Step 2: Which methods are suitable to correct for the response styles in the data?

Because the analysis in step 1 suggests that both ARS and ERS are likely to be present in the data, derived data sets were constructed using the following two correction methods: (1) substraction of individual means (to account for ARS) and (2) substraction of individual means and division by the individual standard deviations (to account for both ERS and ARS). These corrections are made on an individual and a group-wise level. Group-wise corrections were performed using citizenship as an indicator for a similar cultural background.

Step 3: Do all data sets lead to the same conclusion about student evaluations?

Three different comparisons were computed: (1) a marketing instructor teaching a postgraduate subject compared with an instructor from another faculty teaching a postgraduate subject, (2) a marketing instructor teaching a postgraduate subject compared with an instructor from the same faculty, but another discipline (accounting) teaching a postgraduate subject and (3) a marketing instructor teaching a postgraduate subject compared with another marketing instructor teaching an undergraduate subject.

A marketing instructor compared with an education instructor

The first subject (Marketing Management) was evaluated by 14 students. Eleven students (79 per cent) were international students, predominantly Asian. The second subject (Models of Behaviour Management) was evaluated by 12 students. All twelve students were local students from Australia.

Table 3 gives the results and contains the estimated mean difference (and the corresponding p-value for the Welch *t*-test in parentheses) for each data set: the raw data and the four derived data sets. Positive differences indicate that the evaluation of the second instructor (Models of Behaviour Management) was better than the evaluation of the first instructor; negative differences indicate that the first instructor (Marketing Management) achieved superior results.

For ease of interpretation, grey shadings indicate differences which are significant at the five per cent level. In cases where the entire row of Table 3 is grey and the signs of the coefficients are identical, subject evaluations unambiguously differ, and can therefore be interpreted as valid differences across subjects. In cases where the entire row in Table 3 is white, no significant difference between subjects exist, and this conclusion (again) can be drawn safely because the various derived data set results confirm the raw data sets results. If a row in Table 3 contains both grey and white cells, the analysis of raw and derived data sets lead to contradictory results. In this case, conclusions about the comparative evaluation of the two subjects cannot be safely drawn. Further data collection would be required to be able to make a final decision.

---------- please insert Table 3 approximately here ----------

Inspection of Table 3 leads to the following insights: based on the raw data, the instructor from the Faculty of Education outperforms the marketing instructor in two dimensions, namely: helping students to think critically/analytically and overall satisfaction with the subject. These results would disadvantage the marketing instructor in a promotion process — or at least give the impression that improvement in these two dimensions is required.

Accounting for response styles leads to a different conclusion: no difference in the teaching performance of the two instructors can be identified. In the case of one single data set (the one assuming the existence of ARS at individual level), it is the marketing lecturer who performs better by clarifying the learning objectives for the subject well. However, this occurs only for one data set, and should therefore not be interpreted because (as mentioned before) it is not known whether it is indeed the individual-level ARS that biases the data. Based on these results, the Faculty of Education and marketing lecturer must be assessed as equally good teachers.

A marketing instructor compared with an accounting instructor

The first subject is again Marketing Management, which was evaluated by 14 students and had 79 per cent international students, predominantly Asian. The second subject (Professional Practice — Taxation) was evaluated by 22 students. Only one student (five per cent) was a local student, whereas the remaining 21 students were internationals students, predominantly Asian.

Table 4 provides the results, and contains the estimated mean difference (and the corresponding p-value for the Welch *t*-test in parentheses) for each data set: the raw data and the

four derived data sets. Positive differences indicate that the evaluation of the second instructor (Professional Practice — Taxation) was better than the evaluation of the first instructor; negative differences indicate that the first instructor (Marketing Management) achieved superior results.

---------- please insert Table 4 approximately here ----------

Inspection of Table 4 leads to the following insights: based on the raw data, the instructors from the Schools of Marketing and Accounting performed equally well as teachers of their respective subjects (no differences in scores were statistically significant). However, when response styles are accounted for, the picture changes: the marketing instructor performs better in several dimensions. With respect to "clarifying the learning objectives", all four derived data sets indicate that the marketing instructor performed better. Evidence also exists that the marketing instructor was better in helping students to develop a good understanding of the content of the subject, although the data set that assumes a group-wise ARS and ERS bias was marginally insignificant. Only a minority of data sets lead to the conclusion that the marketing instructor was also better in helping student to think critically/analytically and was given higher overall satisfaction ratings. From the analysis of the full range of data sets, which account for the existence of response styles in the data, it therefore must be concluded that the marketing instructor performed better in at least one dimension compared to the Accounting lecturer.

A postgraduate marketing instructor compared with an undergraduate marketing instructor

The first subject is again Marketing Management, which was evaluated by 14 students and had 79 per cent international students, predominantly Asian. The second subject (Marketing Communications and Advertising) was evaluated by 20 students, 17 of whom (85 per cent) were Australian. The first subject was a postgraduate subject, that is, it was taken by students following a master's degree in their first year. The second subject was a Level 3 undergraduate subject, that is, it was taken by students following a bachelor's degree in their third year.

Table 5 contains the estimated mean difference (and the corresponding p-value for the Welch *t*-test in parentheses) for each data set: the raw data and the four derived data sets. Positive differences indicate that the evaluation of the second instructor (Marketing Communications and Advertising) was better than the evaluation of the first instructor; negative differences indicate that the first instructor (Marketing Management) achieved superior results.

---------- please insert Table 5 approximately here ----------

Inspection of Table 5 leads to the following insights: based on the raw data, the conclusion would be drawn that the instructor teaching Marketing Management has performed better on four of seven evaluation criteria (clear learning objectives, good understanding of content, critical thinking and enthusiasm to learn more). These results would disadvantage the instructor of the Marketing Communication and Advertising subject in a promotion process.

Accounting for response styles changes the interpretation dramatically: there was no agreement on differences between the two instructors across derived data sets. These results should be interpreted as both instructors performing equally well as teachers in the respective subjects. This can be relatively safely assumed for five of the seven questions, and for the remaining two questions, the majority of computations still suggest insignificant differences. These results could not form a valid basis for a negative evaluation of the instructor of the Marketing Communication and Advertising subject in a promotion process.

CONCLUSION

Student evaluation results are affected by cross-cultural response styles. Ignoring the biasing effect of cross-cultural response styles when comparatively assessing the teaching performance of instructors can lead to biased evaluations. Such biased evaluations can have serious negative consequences for both declared aims of student evaluations: teachers may receive misleading feedback about the areas where they have potential for improvement, and university administration may make misguided decisions regarding tenure and promotion.

These findings have major practical implications for universities that use survey-based student evaluations in general and marketing instructors in particular:

1. Universities that have a diverse mix of students from different countries of origin and do not routinely include country of origin in their evaluation surveys should do so. If evaluation data is collected without any background information on the cultural background it is impossible to check whether the data is contaminated by cross-cultural response styles. Because business

faculties traditionally have higher proportions of international students, it is more likely that they are disadvantaged in comparative studies of student evaluations.

2. Members of promotion and tenure committees should be educated about the dependence of student evaluations on external factors that are not under the control of the instructor. This includes the faculty in which the evaluation is undertaken and the number of students in the lecture (both factors which frequently are accounted for by universities), as well as student composition in terms of cultural background and the effects different compositions have on evaluation results. Business faculties should push for more education and information for committee members, because their staff are most likely to be disadvantaged by comparisons that do not account for cross-cultural response bias.

3. Universities should regularly validate evaluation data and assess whether any other external factors significantly affect evaluation outcomes, as well as assessing the robustness of different items contained in the questionnaire to external influences.

4. Universities should regularly update their evaluation surveys to account for the findings from their validity analyses. For example, questions found to depend heavily on external factors should be eliminated.

5. Universities should consider using answer formats that are less prone than multi-category scales to capture response styles, such as binary answer formats or even best-to-worst scaling, where students can assess any given instructor relative to other instructors who have taught them in their degree, or an average thereof. Alternatively, marketing departments could develop their own (valid and culturally robust) evaluation tools and use them as a basis for

improvement and as additional evidence for promotion and tenure committees to address points 3, 4 and 5 if improvement cannot be stimulated at university level.

6. Because it is unlikely that universities will change their approaches quickly, marketing instructors assessed on the basis of teaching evaluations should include in their cases (for promotion, tenure or annual review) the details of the subjects they have taught, including the student mix with respect to cultural background if they have reason to assume that the student mix disadvantages them due to cross-cultural response styles.

APPENDIX — Technical details of the artificial data set

Modelling of response styles

The underlying true evaluations which students have of instructors were assumed to be continuous/metric in nature. However, the answer format offered to students in the questionnaire was ordinal, offering them only six response options. This means that students must "translate" their feeling onto the six-point scale provided in the questionnaire. How students translate their feelings differ, depending on their individual response styles; that is, the breakpoints that induce the values of the latent variable to correspond to each category are varied. This approach is in line with similar models which account for response styles proposed by Rossi et al. (2001), Wolfe and Firth (2002), Johnson (2003) and Javaras and Ripley (2007).

The breakpoints of students' "translation function" between the latent continuous true evaluation and the measurement on the six-point scale are determined using equidistant quantiles of a Gaussian distribution. However, the mean and standard deviation of the Gaussian distribution are varied in order to model ARS and ERS. The lower the mean of the distribution, the higher is the contamination with ARS as the breakpoints are shifted to the left; that is, the same values on the latent variable will be mapped to higher categories. Similarly, by reducing the spread, the range of values of the latent variable mapped to the endpoints is increased. Local students are assumed to have a slight tendency for the extreme responses (ERS). This is operationalised by using the quantiles of a Gaussian distribution with mean 0 (zero) and standard deviation 1 (one) as breakpoints, which implies that the extreme answer options are the most frequently ticked. International students are assumed to have a tendency to agree, and exhibit a

mild response style which is modelled by generating their answers using the quantiles of a Gaussian distribution with mean -0.5 and standard deviation 2 as breakpoints.

Modeling of the true teaching quality evaluations by students

To allow for small differences between individuals and random noise, the latent continuous evaluations of students were sampled independently from Gaussian distributions, with a standard deviation equal to 1 (one) and means which are assumed to be the same for each course and question. For the comparison of the two courses, the mean evaluations are assumed to be the same for the first three questions (equal to -1, 0 and 1), while the next two questions have a better evaluation in the second course (equal to -1 and 0 for the first and 0 and 1 for the second course). The last two questions have a better evaluation in the first course (equal to 0 and 1 for the first and -1 and 0 for the second course). Evaluations of three different courses (two in addition to the one used for comparison) are assumed to be available for each student, and the total information for each student is used to correct for response styles. For the remaining two course evaluations (without the one used for comparison), the means of the latent continuous evaluations are randomly sampled from a standard Gaussian distribution.

REFERENCES

Appleton-Knapp, S. L. & Krentler, K. A. (2006). Measuring student expectations and their effects on satisfaction: The importance of managing student expectations. *Journal of Marketing Education,* 28(3), 254–264.

Babin, L. A., Shaffer, T. R. & Tomas, A. M. (2002). Teaching portfolios: Uses and development. *Journal of Marketing Education,* 24(1), 35–42.

Baumgartner, H. & Steenkamp, J. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research,* 38(2), 143–156.

Bertsch, T. & Peek, L. (1982). Determination of measurement scales for revising or developing teacher evaluation instruments. *Journal of Marketing Education,* 4(1), 15–24.

Byrne, B. M. & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure — A look beneath the surface. *Journal of Cross-Cultural Psychology,* 30(5), 555–574.

Cheung, G. W. & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modelling. *Journal of Cross-Cultural Psychology,* 31(2), 187–212.

Chun, K. T., Campbell, J. B. & Yoo, J. H. (1974). Extreme response style in cross-cultural research — Reminder. *Journal of Cross-Cultural Psychology,* 5(4), 465–480.

Clarke III, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review,* 18(3), 301–324.

Clarke III, I. & Flaherty, T. B. (2002). Teaching internationally: Matching part-time MBA instructional tools to host country student preferences. *Journal of Marketing Education*, 24(3), 233–242.

Clayson, D. E. (1999). Students' evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education,* 21(1), 68–75.

Clayson, D. E. & Haley, D. A. (1990). Student evaluations in marketing: What is actually being measured? *Journal of Marketing Education,* 12(1), 9–17.

Clayson, D. E. & Sheffret, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education,* 28(2), 149–160.

Cronbach, L. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement,* 10, 3–31.

De Jong, M.G., Steenkamp, J.-B. E.M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, 45(1), 104-115.

Dolnicar, S. & Grün, B. 2007a. Assessing analytical robustness in cross-cultural comparisons. *International Journal of Culture, Tourism and Hospitality Research,* 1(2), 140–160.

Dolnicar, S. & Grün, B. 2007b. Cross-cultural differences in survey response patterns. *International Marketing Review,* 24(2), 127–143.

Dolnicar, S., Grün, B. & Le, H. (In press). Cross-cultural comparisons of tourist satisfaction: Assessing analytical robustness. In A. Yuksel (Eds.), *Tourist Satisfaction and Complaining*

Behaviors: Measurement and Management Issues in the Tourism and Hospitality Industry. US: Nova Science Publishers.

Fischer, R. (2004). Standardisation to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology,* 35(3), 263–282.

Greenleaf, E. A. (1992a). Improving rating-scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research,* 29(2), 176–188.

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly,* 56(3), 328–351.

Grunenwald, J. P. & Ackerman, L. (1986). A modified Delphi approach to the development of student evaluations of faculty teaching. *Journal of Marketing Education,* 8(2), 32–38.

Heide, M. & Gronhaug, K. (1992). The impact of response styles in surveys: A simulation study. *Journal of the Market Research Society,* 34(3), 215–223.

Hui, C. H. & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology,* 20(3), 296–309.

Javaras, K.N. & Ripley, B.D. (2007). An "unfolding" latent variable model for Likert attitude data: drawing inferences adjusted for response styles. *Journal of the American Statistical Association*, 102, 454-463.

Johnson, T.R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response styles. *Psychometrika*, 68(4), 563-583.

Lee, J. A., Soutar, G. N., Louviere, J. & Daly, T. M. (2006). An examination of the relationship between values and holiday benefits across cultures using ratings scales and best-worst scaling. ANZMAC CD Proceedings.

Marin, G., Gamba, R. J. & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics — The role of acculturation and education. *Journal of Cross-Cultural Psychology,* 23(4), 498–509.

Marsh, H. W., Hau, K. T., Chung, C. M. & Siu, T. L. P. (1997). Students' evaluations of university teaching: Chinese version of the students' evaluations of educational quality (SEEQ) instrument. *Journal of Educational Psychology, 89* , 568-572.

Marsh, H. W., Hau, K. T., Chung, C. M., & Siu, T. L. P. (1998). Confirmatory factor analysis of students' evaluations: Chinese SEEQ version. *Structural Equation Modeling*, 5, 143-164.

Marshall, R. & Lee, C. (1998). A cross-cultural, between-gender study of extreme response style. *European Advances in Consumer Research,* 3, 90–95.

Orsini, J. L. (1988). Halo effects in students evaluations of faculty: A case application. *Journal of Marketing Education,* 10(1), 38–45.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). San Diego: Academic Press.

Rossi, P. E., Gilula, Z. & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association,* 96(453), 20–31.

Simpson, P. M. & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education,* 22(3), 199–213.

Spivey, W. A. & Caldwell, D. F. (1982). Improving MBA teaching evaluations?: Insights from critical incidents methodology. *Journal of Marketing Education,* 4(1), 25–30.

van de Vijver, F. J. R. & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology,* 33(2), 141–156.

van Herk, H., Poortinga, Y. H. & Verhallen, T. M. M. (2004). Response styles in rating scales — Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 35(3), 346–360.

Welkenhuysen-Gybels, J., Billiet, J. & Cambre, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology,* 34(6), 702–722.

Wheeler, G. E. & Geurts, M. D. (1986). Student evaluations in faculty: A longitudinal study from one department in a business school. *Journal of Marketing Education,* 8(1), 32–38.

Wilson, T. C. (1982). Student friendly evaluation forms in marketing: A review. *Journal of Marketing Education,* 4(1), 7–14.

Wolfe, R. & Firth, D. (2002). Modelling subjective use of an ordinal response scale in a many period crossover experiment. *Applied Statistics*, 51, 245-255.

Zax, M. & Takahashi, S. (1967). Cultural influences on response style: Comparison of Japanese and American college students. *Journal of Social Psychology,* 71, 2–10.

TABLES AND FIGURES

TABLE 1

PERCENTAGE OF CORRECT RESULTS FOR ARITIFICIAL DATA

Scenario	Raw	Individual		Group-wise	
		ARS	ARS & ERS	ARS	ARS & ERS
Only local students in both courses	99	98	98	99	99
Only international students in both courses	98	98	98	98	98
Half local, half international in both courses	98	98	98	98	98
Once course only local students, the other course only international students	71	85	97	84	97

TABLE 2

FREQUENCIES (RELATIVE PERCENTAGE) OF USING AVAILABLE ANSWER OPTIONS

Country of origin		Strongly agree	Agree	Slightly agree	Slightly disagree	Disagree	Strongly disagree
Local		10,048	15,087	7,225	2,584	1,986	1,437
		(26%)	(39%)	(19%)	(7%)	(5%)	(4%)
International	all	2,250	4,187	1,937	585	367	215
		(24%)	(44%)	(20%)	(6%)	(4%)	(2%)
	only Chinese	1,143	2,003	932	244	137	77
		(25%)	(44%)	(21%)	(5%)	(3%)	(2%)

TABLE 3

ESTIMATED DIFFERENCES (P-VALUES) FOR THE COMPARISONS OF A MARKETING SUBJECT WITH A SUBJECT FROM ANOTHER FACULTY USING THE RAW AND THE CORRECTED DATA.

	Correction method				
	Raw	Individual		Group-wise	
Questions		ARS	ARS & ERS	ARS	ARS & ERS
The learning objectives for this subject were made clear.	0.07 (0.41)	-0.20 (0.02)	-0.31 (0.14)	-0.02 (0.84)	-0.07 (0.76)
The criteria for assessment in this subject were made clear.	0.12 (0.41)	-0.15 (0.25)	-0.03 (0.93)	0.04 (0.80)	0.24 (0.53)
I have developed a good understanding of the content of this subject.	0.18 (0.07)	-0.09 (0.32)	0.10 (0.73)	0.10 (0.34)	0.26 (0.32)
My learning in this subject was well supported.	0.21 (0.12)	-0.06 (0.61)	-0.40 (0.37)	0.13 (0.33)	0.37 (0.24)
This subject helped me to think critically/analytically.	0.20 (0.05)	-0.07 (0.35)	-0.04 (0.85)	0.12 (0.22)	0.32 (0.16)
As a result of my experience with this subject I am enthusiastic about further learning.	0.19 (0.29)	-0.09 (0.48)	-0.21 (0.56)	0.10 (0.54)	0.25 (0.52)
Overall, I am satisfied with my learning experience in this subject.	0.22 (0.04)	-0.05 (0.56)	-0.30 (0.42)	0.14 (0.22)	0.40 (0.14)

TABLE 4

ESTIMATED DIFFERENCES (P-VALUES) FOR THE COMPARISONS OF A MARKETING SUBJECT WITH A SUBJECT FROM THE SAME FACULTY BUT ANOTHER SCHOOL USING THE RAW AND THE CORRECTED DATA.

	Correction method				
	Raw	Individual		Group-wise	
Questions		ARS	ARS & ERS	ARS	ARS & ERS
The learning objectives for this subject were made clear.	-0.29 (0.05)	-0.27 (< 0.01)	-0.62 (0.01)	-0.36 (0.02)	-0.80 (0.03)
The criteria for assessment in this subject were made clear.	-0.03 (0.83)	-0.02 (0.86)	0.34 (0.27)	-0.10 (0.52)	-0.07 (0.86)
I have developed a good understanding of the content of this subject.	-0.29 (0.06)	-0.27 (< 0.01)	-0.64 (0.03)	-0.36 (0.02)	-0.75 (0.06)
My learning in this subject was well supported.	-0.14 (0.40)	-0.12 (0.19)	-0.27 (0.30)	-0.21 (0.20)	-0.45 (0.25)
This subject helped me to think critically/analytically.	-0.18 (0.23)	-0.17 (0.02)	-0.31 (0.15)	-0.25 (0.10)	-0.48 (0.20)
As a result of my experience with this subject I am enthusiastic about further learning.	-0.04 (0.80)	-0.03 (0.77)	-0.06 (0.82)	-0.11 (0.47)	-0.27 (0.46)
Overall, I am satisfied with my learning experience in this subject.	-0.16 (0.25)	-0.14 (0.01)	-0.39 (0.04)	-0.23 (0.10)	-0.51 (0.13)

TABLE 5

ESTIMATED DIFFERENCES (P-VALUES) FOR THE COMPARISONS OF TWO DIFFERENT MARKETING SUBJECTS USING THE RAW AND THE CORRECTED DATA.

	Correction method				
	Raw	Individual		Group-wise	
Questions		ARS	ARS & ERS	ARS	ARS & ERS
The learning objectives for this subject were made clear.	-0.24 (<0.01)	-0.15 (0.13)	-0.42 (0.07)	-0.16 (0.06)	-0.39 (0.04)
The criteria for assessment in this subject were made clear.	-0.28 (0.06)	-0.19 (0.18)	-0.14 (0.68)	-0.20 (0.18)	-0.31 (0.37)
I have developed a good understanding of the content of this subject.	-0.27 (0.01)	-0.18 (0.06)	-0.41 (0.07)	-0.18 (0.08)	-0.40 (0.11)
My learning in this subject was well supported.	-0.21 (0.21)	-0.12 (0.41)	-0.09 (0.77)	-0.12 (0.42)	-0.21 (0.53)
This subject helped me to think critically/analytically.	-0.30 (0.04)	-0.21 (0.10)	-0.46 (0.12)	-0.22 (0.14)	-0.44 (0.19)
As a result of my experience with this subject I am enthusiastic about further learning.	-0.33 (0.04)	-0.25 (0.08)	-0.63 (0.04)	-0.25 (0.10)	-0.55 (0.11)
Overall, I am satisfied with my learning experience in this subject.	-0.20 (0.13)	-0.11 (0.31)	-0.30 (0.29)	-0.12 (0.36)	-0.21 (0.48)

FIGURE 1

ILLUSTRATION OF STEPS 2 AND 3