European Journal of Open Education and E-learning Studies
ISSN: 2501-9120
ISSN-L: 2501-9120
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.1456188 Volume 3 | Issue 2 | 2018

THE BEST-ACHIEVING ONLINE STUDENTS ARE OVERREPRESENTED IN COURSE RATINGS

Ricardo Tejeiro[1i],
Alexander Whitelock-Wainwright[2],
Alina Perez[2],
Miguel Angel Urbina-Garcia[3]

[1]School of Psychology, The University of Liverpool,
Eleanor Rathbone Building, South Bedford Street,
L69 7ZA, Liverpool, United Kingdom
[2]Laureate International Universities,
650 S. Exeter Street, Baltimore,
Maryland 21202-4382, United Strates of America
[3]School of Education and Social Sciences,
University of Hull, Cottingham Road,
HU6 7RX, Hull, United Kingdom

Abstract:
Student ratings are the most used and influential measure of performance in Higher Education, and an integral component of formative and summative decision making. This may be particularly relevant in the relatively new online courses, where the pedagogical model is still developing. However, student ratings face strong controversy, and some remarkable challenges –one of which stems from the fact that not all students provide ratings. Nonresponse bias, or the lack of representativeness of the providers of ratings, has been measured and discussed in traditional courses, but to date no study has analysed nonresponse bias in the online evaluation of a fully online higher education course. Our study aims to close this gap. We analysed archival data for the students completing the intake module of four psychology online postgraduate programmes in a 2-year period (June 2014 to May 2016; n = 457). Statistical analyses included correlation, chi-square test, Mantel-Haenszel test of trend, Mann-Whitney's U and regression analysis; effect size was measured with odds ratios, Cramer´s V, and r. We found that the likelihood of providing ratings was not associated with sex, age, educational background, or familiarity with the British higher education system; however, respondents presented significantly higher values than nonrespondents in the key variable used to measure their learning experience –final mark. The implications of

[i] Correspondence: email Corresponding author: Ricardo Tejeiro, School of Psychology, The University of Liverpool, Eleanor Rathbone Building, South Bedford Street, L69 7ZA, Liverpool (United Kingdom).
E-mail: r.tejeiro@liverpool.ac.uk

this finding are discussed in relation to Groves' (2006) causal models for nonresponse bias, as well as the validity and leniency hypotheses.

Keywords: student ratings; learning analytics; teaching quality; nonresponse; online education

1. Introduction

Teaching and course evaluations are nowadays conducted in almost every college or university (Haladyna & Amrein-Beardsley 2009) within their efforts to provide a high quality service to their students. Evaluations have become an integral component of both formative decision making (e.g., re-designing courses to improve the student experience) and summative decision making, with potential impact on aspects of the academics' career, such as promotion, salary (Avery, Bryant, Mathios, Kang, & Bell 2006; Mau & Opengart 2012), or course assignment (Adams & Umbach 2012).

Whilst there are numerous strategies to measure teaching effectiveness –including peer ratings, self-evaluation, employer ratings, teaching awards, learning outcome measures, teaching portfolios, and others (Berk 2005)–, student ratings (SR) are by far the most used and influential measure of performance (Emery, Kramer, & Tian 2003). In most higher education (HE) institutions, student ratings are collected through a short, standardised questionnaire (Murray 2005), on which students are required to rate their experience through various rating scales, sometimes complemented with free-text notes. Depending on the programme and objectives, questions may refer to the instructor (e.g., expertise, style, feedback), to the course (e.g., structure, textbook, supplementary reading), to student support (e.g., availability of information, management of incidences), and others (e.g., equipment quality, student union perceptions).

However, although most HE teaching staff seem to support SR (Murray 2005), strong opinions both for and against their use persist (Theall & Franklin 2001). On the one hand, some teachers and administrators consider that SR provides a unique and very valuable insight into the teaching-learning process (Gaillard, Mitchell, & Kavota 2006). On the other hand, some staff question the ability of students to make adequate judgments based either in their lack of knowledge about the most appropriate teaching methods (Gravestock & Gregor-Greenleaf 2008), or in their application of spurious evaluation criteria (Steiner, Holley, Gerdes, & Campbell 2006) –e.g., giving higher marks to nicer rather than to the most effective teachers–, in what Gomez-Mejia and Balkin (1992) compare to a popularity contest. Also, it is argued that some students expect little consequences from their comments and therefore provide superficial evaluations (Gaillard et al. 2006).

Not surprisingly, SR generate a high volume of research. Regarding reliability, findings suggest that "*ratings of a given instructor are reasonably stable or consistent across courses, years, rating forms, and groups of raters, [and that SR] agree with evaluations made by others, such as colleagues and alumni*" (Murray 2005, p.2). However, results about their

validity – the other great issue around SR (Donovan, Mader, & Shinsky 2007; Pritchard & Potter 2011) – seem to be less consistent. The question here is –do SR really measure teaching quality?

Some studies have addressed this issue by analysing the variables that affect the valence of SR, this is, what makes students provide more (or less) positive ratings. These include instructor characteristics such as organisation, clarity, availability and knowledge of the subject (Marsh & Roche, 1999), classroom characteristics like time of day, class size, subject matter and class level (McPherson 2006; Millea & Grimes 2002), characteristics of the course such elective vs. required (Darby 2006), and online vs. offline delivery (Ardalan, Ardalan, Coppage, & Crouch 2007; Avery et al. 2006; Donovan et al. 2007; Mau & Opengart 2012).

One of the most interesting findings in this regard is the repeatedly documented significant correlation between expected mark and valence of SR; in other words, students with higher expected marks tend to assign higher marks to their instructors (Blackhart, Peruche, DeWall, & Joiner 2006; Crumbley & Reichelt 2009; Denson, Loveday, & Dalton 2010; Heine & Maddox 2009; Marsh 2007). There is no consensus regarding the interpretation of this correlation, but two main proposals are often cited. The *validity explanation* argues that higher marks are in fact the result of a better teaching (Heckert, Latier, Ringwald, & Drazen 2006), whereas the *leniency hypothesis* suggests that students give higher scores to instructors from whom they receive, or expect to receive good marks (McPherson 2006); an obvious consequence would be that instructors are dangerously tempted to artificially raise marks (Pritchard & Potter 2011). Additionally, some studies suggest that the correlation between SR and measures of student learning may be limited to some forms of assessment (Stehle, Spinath, & Kadmon 2012).

1.1 Nonresponse bias

A different line of research stems from the fact that usually not all the students who form the target population do provide ratings –this is, there is some level of nonresponse rate. This may be due to a variety of reasons, such as *inaccessibility* (the student did not receive the rating form), *inability* to respond (due to illness, temporary interruptions in internet service and other causes), *carelessness* (the student lost the survey or exceeded the deadline) and *noncompliance* (they decided not to participate) (Sosdian & Sharp, 1980).

It may be the case that students who provide SR are systematically different to those who do not respond in variables relevant to the evaluation. This results in a measurement error known as nonresponse bias (McDaniel & Gates 2012; Sax, Gilmartin, & Bryant 2003). As Ellis, Endo and Armer (1970) systematize it, nonresponse bias (NRB) is "*a function of: (a) the proportion of nonrespondents in the total sample and (b) the extent to which there is a systematic discrepancy between respondents and nonrespondents on variables relevant to the inquiry*" (p. 103). From this definition, it is evident that low response rates, whilst increasing the possibility of error bias (Adams & Umbach 2012; Groves &

Peytcheva 2008; Porter & Umbach 2006), will only result in it when respondents' and nonrespondents' characteristics are actually different (Dillman 1991).

NRB poses a substantial threat to the accuracy of results (McDaniel & Gates 2012) and to the external validity of conclusions (Micklewright, Schnepf, & Skinner 2012). Wrong decisions are then likely to be taken, and the fear that students who complete SR are not representative of all students in the class may make teachers disregard the information provided by these tools (Sax et al. 2003). A common approach to minimise the impact of NRB, is to set a minimum threshold that response rates must exceed for the survey to be acceptable – different levels have been suggested, including 75% (Ary, Jacobs, & Razavieh 1996), 80% (Gall, Borg, & Gall 1996; Tuckman 1999), 85% (Lindner et al. 2001), and 90% (Miller & Smith 1983). However, NRB is not only a function of the proportion of nonrespondents in the sample: it does not occur if respondents and nonrespondents do not differ substantially, even when the response rate is very low.

The recognition of how NRB can be detrimental to a comprehensive understanding of student perceptions does motivate efforts to identify, control and when possible minimise the sources of nonresponse. Gender is commonly cited as one of the main socio-demographic sources of variation, with females usually found to respond more than males (Avery et al. 2006; Donovan et al. 2007; McInnis 2006; Porter & Umbach 2006; Porter & Whitcomb 2005; Reisenwitz 2016; Sax et al. 2003; Sax, Gilmartin, Lee, & Hagedorn 2008). In fact, it has been reported that female students are more serious in the evaluation process and consider it as more important than male students (Heine & Maddox 2009). Higher likelihood to provide SR has also been reported for high-achieving students (Adams & Umbach, 2012; Avery et al. 2006; Porter and Umbach, 2006; Porter & Whitcomb, 2005).

Other factors that may affect nonresponse include ethnicity (Porter & Umbach 2006), personality (Marcus & Schutz 2005; Porter & Whitcomb 2005), attitudinal characteristics (Hochstim & Athanasopoulos 1970), technology savviness (Reisenwitz 2016), salience of the survey (Adams & Umbach 2012; Groves et al. 2006), fear of being identified (Olsen 2008), perceived lack of skills to provide constructive feedback (Nulty 2008), survey fatigue/saturation (Adams & Umbach 2012; Gee 2015), timing of the evaluation (Estelami 2015), and institutional factors such as urban location, public status, and an increased density of students (Porter & Umbach 2006).

Mode of delivery of the evaluations has also been identified as a key factor, with researchers systematically concluding that the percentage of students completing online evaluations is smaller than the percentage of those who complete the evaluations in class (Ardalan et al. 2007; Avery et al. 2006; Guder & Malliaris 2013; Nowell, Gale, & Handley 2010; Nulty 2008). Online evaluations of teaching are in fact more and more common, as the use of the internet and of virtual learning environments expands across most if not all universities. In parallel, the number of studies specifically analyzing the factors that influence online SR –including nonresponse– have increased in the past decade. For instance, Adams and Umbach (2012) studied the influence of salience, fatigue, and academic environments on nonresponse rates; Reisenwitz (2016) analysed the demographic variables that contribute to nonresponse bias in online student

evaluations; and several authors have compared online versus traditional pen-and-paper evaluations (Donovan et al. 2007; Liegle & McDonald 2005; Nowell, Gale, & Kerkvliet 2014; Sax et al. 2003). However, whilst all these studies have analysed the use of online SR for on-campus courses, little is still known about the online evaluation of online HE courses. If, as previously stated, NRB can lead to mistakes in both formative and summative decision making, this might be even more relevant in fully online programmes, where the availability of other sources of feedback (e.g., frequent face-to-face discussions with students and colleagues, or informal observation of students' behaviour) is substantially reduced, and where the evidence to support teaching strategies and methods is still limited.

The aim of this study is to contribute to closing this gap by analyzing some of the factors associated with nonresponse in the online SR of a module corresponding to a fully online postgraduate programme delivered by a major HE institution in the United Kingdom. Also, we intend to analyse not only two of the most commonly studied factors (students' sex and marks), but also age and familiarity with the evaluation context –both largely neglected in the studies. Specifically, from the literature above we hypothesize that response will be higher amongst (1) females as compared to males, and (2) students with higher marks as compared to students with lower marks. Considering Groves, Presser & Dipko's (2004) finding that participation in surveys is higher when the topic is of interest to participants, as well as Groves and Petycheva's (2006) finding that NRB was lower when the sample had prior involvement with the survey sponsor, we hypothesize that (3) greater familiarity with the context of the evaluation will result in a higher response rate. The relationship between age and response will be analysed in an exploratory manner.

1.2 Background of Institution

This study was conducted at a public research university in the North of England, which has offered postgraduate psychology online programmes since 2012. All students completing each taught 8- or 12-week module are invited to complete, during the last week of the module and the next one after it, a standardized 25-question survey where they rate, on a 5-point scale ranging from 1 to 5, several aspects of their instructor's teaching (satisfaction, expertise, style, feedback, etc.), and of the module structure, consistency, navigation, textbook, and workload, amongst other aspects. Participation is voluntary and students are informed that their answers will remain anonymous, although access to the survey requires the students' regular username and password.

2. Method

2.1 Sample

After obtaining the approval of the University's ethics committee (reference 0871), data were collected from the 457 students who completed the intake module (common to the four postgraduate programmes in psychology) between June 2014 and May 2016; the first date corresponds to the moment when the current programmes were launched,

and the second was selected to complete a 2-year period. However, for the 13 students who failed the module and had to retake it during this period, only the data for the second intake were considered. Additionally, due to administrative or technical reasons, no data on survey completion were available for 21 students and no personal data were available for one student; these students were removed from the sample.

The final sample was thus formed by 421 students, 66.3% of them females (n = 279), aged 21 to 64 years (Median = 35); they had 86 different nationalities (23.5% were Britons, 7.6% were Canadians, 5.2% were South-Africans, and the remaining had a presence below 5%) and resided in 81 different countries (21.6% in the UK, 7.8% in Canada, 7.6% in the United Arab Emirates, and the remaining were very distributed). They started the module in 30 classes across 4-5 intakes per year, with 2-4 classes per intake. There were between 10 and 20 students per class (M = 14.03, SD = 3.02).

2.2 Analyses

Two different levels of analyses were used. First, we analysed the data at class level: the end-of-module reports were used to retrieve measures of success for each class (percentage of successful completion, percentage of marks above pass level, average score given by the students to the overall module, and average score given to the instructor). The relationship between these variables and response rate was then analysed using Person's and Spearman's correlations (Shapiro-Wilk tests revealed that scores were normal for the percentage variables, with p = .449 and p = .893, respectively; and non-normal for the score variables, with p = .007 and p = .002, respectively).

Second, we analysed the data at module (whole sample) level. We utilised the record-linkage analysis approach, where both respondents and nonrespondents are linked to database records that are available for the full sample (Porter & Whitcomb 2005). Data were collected from two sources. On the one hand, the admission file completed for each student was used to retrieve sex, age, educational background, nationality and country of residence. No data regarding other potentially relevant variables –e.g., socio-economic status or ethnicity– are collected from students and therefore they are not available in their files. The variables were coded for educational background: completion of a Bachelor degree (or its international equivalent), whether or not the degree was in psychology (or had substantial relationship with psychology), and if the individual had personal experience as a student or as a teacher in higher education in the UK. The three variables were dichotomously coded (1 = yes, 0 = no). Independent and blind coding of these variables on a random sample of 20 participants was carried out by the fourth researcher, with 100% coincidence with the main coder (first researcher). Nationality and country of residence were also dichotomously coded (1 = UK, 0 = non-UK). The inclusion of three educational and two country variables intended to measure the familiarity of the student with the rating scenario; these variables were first analysed independently, and then combined into an ordinal variable termed 'familiarity', with a score of 0-5 according to the number of dichotomous IVs with a YES answer. Whilst we acknowledge that familiarity with other HE systems in other parts of the world may have different impact on students'

completion of rating surveys, exploring that possibility was out of the scope of the present study –given the vast array of backgrounds in our sample.

On the other hand, the information in the virtual learning environment (VLE) was used to retrieve, for each student, their final mark in the module and whether or not they had completed the end-of-module survey. It must be noted that the specific responses given by each student to the survey are not available to anyone because feedback data are collated as a whole and no individual responses are stored in the system.

Chi-square tests were used to measure the significance of the differences in response rate between the groups with each of the two values in the nominal variables, with odds ratio as a measure of effect size. The differences in response rate associated with familiarity –measured in an ordinal scale– were analysed using Mantel-Haenszel test of trend, with Cramer´s V (ϕ_c) as a measure of effect size. The differences in age and mark between respondents and nonrespondents were analysed with Mann-Whitney's U (because the two variables were non-normally distributed; Shapiro-Wilk's tests $p < .001$), with r as a measure of effect size. Additionally, logistic regression was used to further analyse the relationship between each independent variable (IV) and the dependent variable (DV), controlling for individual differences in the other IVs.

3. Results

3.1 Class level

Due to technical reasons, the full data required for analyses was only available for 22 out of the 30 classes in the study. Response rates were normally distributed (Shapiro-Wilk's test $p = .981$) and ranged from 24% to 87% ($M = 56.3$, $SD = 0.16$); in terms of quartiles, one class (3.3%) was in the first quartile with a response rate between 0 and 25%, 7 classes (23.3%) were in the second quartile, 11 classes (36.7%) were in the third quartile, and 3 classes (10%) were in the fourth quartile. Response rate correlated positively with all measures of class success, although the relationships failed to reach significance ($r = .29$, $p = .185$ for percentage of success; $r = .31$, $p = .164$ for percentage of marks above pass level; $r_s = .31$, $p = .162$ for average class score; $r_s = .34$, $p = .128$ for average instructor score).

3.2 Module level

Table 1 shows the percentages of response and the significance of differences between each value of the dichotomous variables. Although higher percentages of response corresponded to females over males, those with a Bachelor degree over those without it, those with a degree in psychology slightly over those without it, and students with no previous HE experience in the UK over those with previous experience, all differences were non-significant. Familiarity, measured by a combined ordinal scale, was not related to the likelihood of SR; Mantel-Haenszel $\chi^2(1, n = 421) = 0.12$, $p = .728$, $\phi_c = .09$. Mann-Whitney's tests revealed that respondents and nonrespondents did not differ in age (Mdn = 35 years for respondents and 36 for nonrespondents; $U = 20395$, $p = .630$, $r =$

.02). However, they differed significantly in mark, with respondents (Mdn = 65.38) presenting higher marks than nonrespondents (Mdn = 58.18); U = 10543, p < .001, r = .42.

Table 1: Percentages of response for each value of the dichotomous variables, and significance of differences

Variable	Value	*n*(%)	$\chi_{(1,\ n\ =\ 421)}$	*p*	*OR*
Sex	Males	79(55.6)	3.14	.077	1.45, 95% CI [0.96-2.19]
	Females	180(64.5)			
Bachelor	Yes	231(60.5)	1.92	.166	1.66, 95% CI [0.80-3.44]
	No	28(71.8)			
Psychology	Yes	90(61.6)	0.00	.970	1.01, 95% CI [0.67-1.52]
	No	169(61.5)			
HE UK	Yes	79(60.3)	0.12	.731	1.08, 95% CI [0.71-1.64]
	No	180(62.1)			

It may be the case that very low final marks correspond mostly to individuals who disengaged from the classes before their end, but did not retrieve formally from the module (data for those who retrieved from the module were not available). Disengagement may be due to academic reasons (skills, dissatisfaction with contents or activities, inadequate teaching methods…) but also to other types of reasons (work or financial issues, unexpected personal events, health problems…); the reason for each individual case's disengagement was not available. Including these students and their marks into the database may overestimate the relationship between low marks and nonresponse, as these students are very unlikely to complete the end-of-module survey. For this reason, the analyses above were repeated only on the students who passed the module (n = 347; 69.5% of them females). All results were similar to those obtained on the whole sample.

Confirmation tests were conducted on the whole sample using logistic regression, with response as DV and sex, age, mark and combined familiarity as IVs. Also, all possible IV interaction terms were created and tested. With this procedure, we intended to measure the relationship of each IV with the DV, controlling for the effect of the remaining IVs. This is important because males and females were found to differ significantly in age and mark (Mann-Whitney's U = 16811, p = .011, r = .12 for age; U = 16747.5, p = .009, r = .13 for mark), although not in familiarity (U = 19484, p = .774, r = .01). The only significant effect on the DV was found for mark, with β = 0.075 (1), Wald = 55.03, p < .001, *OR* = 1.08, 95%CI [1.06-1.10]. A one-unit increase in mark was associated with 1.08 times greater likelihood that the student provides rating (i.e., is a respondent). Chi-square was nonsignificant at p = .070 in the Hosmer and Lemeshow test, which indicates that the data fit the model well; also, Nagelkerke's R square was

.339, which roughly implies that 33.9% of the variability in the DV (likelihood of response) is accounted for by the IV (mark).

4. Discussion

To the best of our knowledge, this is the first study on NRB in a fully online programme. Our results only partly resemble those from previous research, and therefore only one of our three hypotheses receives support. Contrary to the findings in most studies –and to our first hypothesis–, females were as likely as males to provide SR, even when the other possible IVs were controlled for. It is possible that the mode of delivery (online), the level of education (postgraduate) or the subject of the programme (psychology) are partly responsible for our results. On the other hand, our results are in consonance with those of Spencer and Schmelkin (2002), who found no meaningful differences between male and female students' attitudes towards SR.

As indicated in our third hypothesis, we also expected that greater familiarity with the context in which the evaluation takes place would promote more positive attitudes which, in turn, would result in a higher response rate. However, we found no association between response rate and any measure of familiarity, either independently considered or in the form of a combined scale. Perhaps greater familiarity only promotes better attitudes when it implies the perception of a positive impact on the students' experience (e.g., via timely and effective responses from the academic institution), which might not necessarily reflect the experience of many students. In fact, it has been reported that the importance given to SR by graduate students and seniors is lower than that given by juniors and sophomores –although no possible explanations for this difference were suggested or explored (Spencer & Schmelkin 2002).

The second hypothesis (students with higher marks will be more likely to provide SR, as compared to students with lower marks), was supported. In fact, mark was the only variable significantly associated with response rate. It must be noted that the final mark in the module is the weighted average of the marks obtained in the different weekly assignments, and that at the moment of completing the survey, some marks have not yet been communicated to the students –although they already know others and can make a rough estimation of their final result.

What do these results tell us about the validity of the students' ratings, as used in our programmes? First: the students who provide SR may not be representative of the population of students who *register* in the module (as those who voluntarily retrieve from it do not provide any feedback, although it seems sensible to assume that at least some of them may have dropped-out due to dissatisfaction with their learning experience). Second, the students who provide SR may not be representative of the population of students who *complete* the course (as the students who complete but fail are under-represented and those who complete and pass are over-represented). Third, the students who provide SR may not be representative of the population of students who *pass* the course (as those who pass with higher marks are more likely to provide SR than those who pass with lower marks). These results do not directly imply that the

course is either over- or under-rated; they simply indicate that those who provide SR are significantly different –in academic success in the module– from those who do not provide SR.

We can now go back to Ellis et al.'s (1970) definition of nonresponse bias as "a function of: (a) the proportion of nonrespondents in the total sample and (b) the extent to which there is a systematic discrepancy between respondents and nonrespondents on variables relevant to the inquiry" (p. 103). With regard to the first criterion, the average percentage of nonrespondents in the classes analysed was 44%, with only one out of ten classes exceeding the 75% participation that constitutes the lowest of the thresholds suggested (Ary et al., 1996), and none reaching the most exigent threshold of 90% (Miller & Smith, 1983). With regard to the second criterion, we found a systematic discrepancy between respondents and nonrespondents in final mark.

The causal models that describe alternative conditions related to NRB, as suggested by Groves (2006), can help in the interpretation of this finding. According to the *separate causes* model, the vector of causes of the Y variable (mark) is independent of the causes of response propensity, P. The *common cause model* asserts that Y (mark) and response propensity have shared causes (Z). Finally, the *survey variable cause model* asserts that Y (mark) is a cause of response propensity.

The existence of covariance between mark and response propensity allows us to reject the separate causes model. With regard to the common cause model, our results exclude gender, age, general academic background or familiarity as shared causes. That causal role might be played by *better learning experience*. As indicated above, previous research has revealed that students with higher expected marks tend to assign higher marks in their ratings (Blackhart et al. 2006; Crumbley & Reichelt 2009; Denson et al. 2010; Heine & Maddox 2009; Marsh 2007). If this is due to better learning experience –as in the validity explanation–, then it may be the case that students who learn more and better are also more engaged in satellite activities such as completing the end-of-module survey. Better learning experience might then be a common cause, Z, for both marks and participation, and it would also explain our finding that participation increased in parallel to class results –although the relationship, with correlation values around .30, failed to reach significance. If better learning experience is behind the covariance between marks and participation, then the validity of the surveys is unchallenged: courses obtain higher scores because they are effectively better. However, *better teaching-learning experience* cannot be directly identified with *better teaching*. It involves a complex process that, rather than a unidirectional influence of the instructor's skills and attitudes, may require an alignment of attitudes, personality traits, needs and interests between teacher and student –further shaped by individual values such as individual responsibility, gratitude, social responsibility, social engagement and even empathy on behalf of the student (see Stronge, Tucker & Hindman 2004).

On the other hand, also the leniency hypothesis can be used to explain our results: more lenient instructors might over-grade some of their students, who in turn may be more likely to participate in the survey (and to provide better scores in it) as a form of reward or compensation. This is consistent with the survey variable cause

model, with marks (Y) being the cause of participation (P). If that is the case, good scores in SR may not necessarily identify good teaching, but perhaps a more lenient one –the scores in the end-of-module surveys are likely to be inflated.

Our study did not have access to other potentially relevant variables that might shed light over the relationships above, such as attitudes, expectations, language, personality, previous experience with feedback, or non-academic events affecting the students' performance –and we used 'proxy' measures for familiarity. In fact, in a sample formed by students from 86 countries, cultural specificities can be expected to play a role in the likelihood to provide feedback –e.g., some students may not feel as comfortable as others when it comes to questioning their instructor's capabilities, skills or knowledge. The potential effect of teachers' instructions was not measured either; e.g., some of them may have played a more active role than others in engaging their students with the survey. Also, we analysed solely psychology postgraduate students, who may differ from students in other levels and disciplines. Additionally, all nonrespondents in our study were analysed together, whilst it has been reported that the various types of nonresponse produce different bias (Campanelli, Sturgis, & Purdon, 1997; Groves & Couper, 1998). Future research should address these limitations and provide an empirically supported interpretation for the relationships identified. In the absence of such an interpretation, it will remain unclear whether the nonresponse bias here identified challenges the validity of student ratings.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Adams, M. J. D., and Umbach, P. D. (2012). Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. *Research in Higher Education,* doi: 10.1007/s11162-011-9240-5

Ardalan, A., Ardalan, R., Coppage, S., and Crouch, W. (2007). A comparison of student feedback obtained through paper-based and web-based surveys of faculty teaching. *British Journal of Educational Technology,* doi: 10.1111/j.1467-8535.2007.00694.x

Ary, D., Jacobs, L., and Razavieh, A. (1996*). Introduction to research in education* (5th Ed.). Ft. Worth, TX: Holt, Rinehar, and Winston, Inc.

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., and Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education,* doi: 10.3200/JECE.37.1.21-37

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48-62.

Blackhart, G. C., Peruche, B. M., DeWall, C. N., and Joiner, T. E. (2006). Factors influencing teaching evaluations in higher education. *Teaching of Psychology, 33*(1), 37–39.

Campanelli, P., Sturgis, P., and Purdon, S. (1997). *Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates*. London: S.C.P.R.

Crumbley, D. L., and Reichelt, K. J. (2009). Teaching effectiveness, impression management, and dysfunctional behavior: Student evaluation of teaching control data. *Quality Assurance in Education,* doi: 10.1108/09684880910992340

Darby, J. A. (2006). The effects of the elective or required status of courses on student evaluations. *Journal of Vocational Education and Training,* doi: 10.1080/13636820500507708

Denson, N., Loveday, T., and Dalton, H. (2010). Student evaluation of courses: what predicts satisfaction. *Higher Education Research and Development,* doi: 10.1080/07294360903394466

Dillman, D. A. (1991). The design and administration of mail surveys. *Annual Review of Sociology,* doi: 10.1146/annurev.so.17.080191.001301

Donovan, J., Mader, C., and Shinsky, J. (2007). Online vs. traditional course evaluation formats: student perceptions. *Journal of Interactive Online Learning, 6*(3), 158-180.

Ellis, R. A., Endo, C. H., and Armer, J. M. (1970). The use of potential nonrespondents for studying nonresponse bias. *Pacific Sociological Review,* doi: 10.2307/1388313

Emery, C. R., Kramer, T. R., and Tian, R. G. (2003). Return to academic standards: A critique of students' evaluations of teaching effectiveness. *Quality Assurance in Education: An International Perspective,* doi: 10.1108/09684880310462074

Estelami, H. (2015). The effects of survey timing on student evaluation of teaching (SET) measures obtained using online surveys. *Journal of Marketing Education,* doi: 10.1177/0273475314552324

Gaillard, F. D., Mitchell, S. P., and Kavota, V. (2006). Students, faculty, and administrators' perception of students' evaluations of faculty in higher education business schools. *Journal of College Teaching and Learning,* doi: 10.19030/tlc.v3i8.1695

Gall, M. D., Borg, W. R., and Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Gee, N. (2015). A study of student completion strategies in a Likert-type course evaluation survey. *Journal of Further and Higher Education*. Advance online publication. doi: 10.1080/0309877X.2015.1100717

Gomez-Mejia, L. R., and Balkin, D. B. (1992). Determinants of Faculty Pay: An Agency Theory Perspective. *Academy of Management Journal, 35*(5), 921-955.

Gravestock, P., and Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends.* Toronto: Higher Education Quality Council of Ontario.

Groves, R. M. and Couper, M. P. (1998). *Nonresponse in Household Interview Surveys.* New York: Wiley.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly,* doi:10.1093/poq/nfl033

Groves, R. M., and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A metaanalysis. *Public Opinion Quarterly,* doi: 10.1093/poq/nfn011

Groves, R. M., Couper, M., Presser, S., Singer, E., Tourangeau, R., Piani Acosta, G., et al. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly,* doi: 10.1093/poq/nfl036

Groves, R. M., Presser, S., and Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly,* doi: 10.1093/poq/nfh002

Guder, F., and Malliaris, M. (2013). Online course evaluations response rates. *American Journal of Business Education, 6*(3), doi: 10.19030/ajbe.v6i3.7813

Haladyna, T., and Amrein-Beardsley, A. (2009). Validation of a research-based student survey of instruction in a college of education. *Educational Assessment, Evaluation and Accountability,* doi:10.1007/s11092-008-9065-8

Heckert, T. M., Latier, A., Ringwald, A., and Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to "buy" better evaluations through lenient grading? *College Student Journal, 40*(3), 588-596.

Heine, P., and Maddox, N. (2009). Student perceptions of the faculty course evaluation process: An exploratory study of gender and class differences. *Research in Higher Education Journal, 3*, 1–10.

Hochstim, J. R., and Athanasopoulos, D. A. (1970). Personal follow-up in a mail survey: Its contribution and its cost. *Public Opinion Quarterly,* doi: 10.1086/267774

Liegle, J., and McDonald, D. S. (2005). Lessons Learned From Online vs. Paper-based Computer Information Students' Evaluation System. *Information Systems Education Journal, 3*(37). Retrieved from http://isedj.org/3/37/. ISSN: 1545-679X.

Lindner, J. R., Murphy, T. H., and Briers, G. H. (2001). Handling nonresponse in social science research. *Journal of Agricultural Education,* doi: 10.5032/jae.2001.04043

Marcus, B., and Schutz, A. (2005). Who are the people reluctant to participate in research? Personality correlates of four different types of nonresponse as inferred from self- and observer ratings. *Journal of Personality,* doi: 10.1111/j.1467-6494.2005.00335.x

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry and J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht: Springer.

Marsh, H. W. and Roche, L. A. (1999). Rely upon SET research. *American Psychologist,* doi: 10.1037/0003-066X.54.7.517

Mau, R. R., and Opengart, R. A. (2012). Comparing ratings: In-class (paper) versus out of class (online) student evaluations. *Higher Education Studies,* doi: 10.5539/hes.v2n3p55

McDaniel, C., Jr., and Gates, R. (2012). *Marketing research* (9th Ed.). Hoboken, NJ: John Wiley.

McInnis, E. D. (2006). *Nonresponse Bias in Student Assessment Surveys: A Comparison of Respondents and Non-Respondents of the National Survey of Student Engagement at an Independent Comprehensive Catholic University* (Doctoral dissertation, Marywood University). Retrieved from http://nsse.indiana.edu/pdf/research_papers/Nonresponse%20Bias%20in%20Student%20Assessment%20Surveys%20-%20Elizabeth%20McInnis.pdf

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *Journal of Economic Education,* doi: 10.3200/JECE.37.1.3-20

Micklewright, J., Schnepf, S. V., and Skinner, C. (2012). Non-response biases in surveys of schoolchildren: the case of the English Programme for International Student Assessment (PISA) samples. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* doi: 10.1111/j.1467-985X.2012.01036.x

Millea, M. and Grimes, P. W. (2002). Grade expectations and student evaluation of teaching. *College Student Journal*, *36*(4), 582–591.

Miller, L. E., and Smith, K. L. (1983). Handling nonresponse issues. *Journal of Extension, 21*(5), 45-50.

Murray, H. G. (2005, June). *Student Evaluation of Teaching: Has It Made a Difference?* Paper presented at the Annual Meeting of the Society for Teaching and Learning in Higher Education, Charlottetown, Canada. Retrieved from https://www.stlhe.ca/wp-content/uploads/2011/07/Student-Evaluation-of-Teaching1.pdf

Nowell, C., Gale, L. R., and Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment and Evaluation in Higher Education,* doi: 10.1080/02602930902862875

Nowell, C., Gale, L. R., and Kerkvliet, J. (2014). Non-response bias in student evaluations of teaching. *International Review of Economics Education,* doi: 10.1016/j.iree.2014.05.002

Nulty, D.D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment and Evaluation in Higher Education,* doi: 10.1080/02602930701293231

Olsen, D. (2008). *Teaching patterns: A pattern language for improving the quality of instruction in higher education settings* (Doctoral dissertation, Utah State University). Retrieved from http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1050andcontext=etd

Porter, S. R., and Umbach, P. D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education,* doi: 10.1007/s11162-005-8887-1

Porter, S. R., and Whitcomb, M. E. (2005). Non-response in student surveys: The role of demographics, engagement, and personality. *Research in Higher Education,* doi: 10.1007/s11162-004-1597-2

Pritchard, R. E., and Potter, G. C. (2011). Adverse changes in faculty behavior resulting from use of student evaluations of teaching: A case study. *Journal of College Teaching and Learning,* doi: 10.19030/tlc.v8i1.980

Reisenwitz, T. H. (2016). Student Evaluation of Teaching: An Investigation of Nonresponse Bias in an Online Context. *Journal of Marketing Education,* doi: 10.1177/0273475315596778

Sax, L. J., Gilmartin, S. K., and Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education,* doi: 10.1023/A:1024232915870

Sax, L. J., Gilmartin, S. K., Lee, J. J., and Hagedorn, L. S. (2008). Using web surveys to reach community college students: An analysis of response rates and response bias. *Community College Journal of Research and Practice,* doi: 10.1080/10668920802000423

Sosdian, C. P., and Sharp, L. M. (1980). Nonresponse in mail surveys: Access failure or respondent resistance. *Public Opinion Quarterly,* doi: 10.1086/268606

Spencer, K. J., and Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment and Evaluation in Higher Education,* doi: 10.1080/0260293022000009285

Stehle, S., Spinath, B. and Kadmon, M. (2012). Measuring Teaching Effectiveness: Correspondence Between Students' Evaluations of Teaching and Different Measures of Student Learning. *Research in Higher Education,* doi: 10.1007/s11162-012-9260-9

Steiner, S., Holley, L. C., Gerdes, K., and Campbell, H. H. (2006). Evaluating teaching: Listening to students while acknowledging bias. *Journal of Social Work Education,* doi: 10.5175/JSWE.2006.200404113

Stronge, J. H., Tucker, P. D., and Hindman, J. L. (2004). *Handbook for Qualities of Effective Teachers.* Alexandria, VA: Association for Supervision and Curriculum Development.

Theall, M., and Franklin, J. L. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P. C., Abrami, and L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them? (New Directions for Institutional Research, No. 109)* (pp. 45–56). San Francisco, CA: Jossey-Bass.

Tuckman, B. W. (1999). *Conducting education research* (5th ed.). Fort Worth, TX: Harcourt Brace.