

Quantifying halo effects in students' evaluation of teaching: a response to Michela

Edmund Cannon & Giam Pietro Cipriani

To cite this article: Edmund Cannon & Giam Pietro Cipriani (2024) Quantifying halo effects in students' evaluation of teaching: a response to Michela, *Assessment & Evaluation in Higher Education*, 49:1, 66-71, DOI: [10.1080/02602938.2023.2180484](https://doi.org/10.1080/02602938.2023.2180484)

To link to this article: <https://doi.org/10.1080/02602938.2023.2180484>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 21 Feb 2023.



Submit your article to this journal [↗](#)



Article views: 799



View related articles [↗](#)





View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)

Quantifying halo effects in students' evaluation of teaching: a response to Michela

Edmund Cannon^a  and Giam Pietro Cipriani^b 

^aSchool of Economics, University of Bristol, Bristol, UK; ^bDepartment of Economics, University of Verona, Verona, Italy

ABSTRACT

In Cannon and Cipriani (2022) we contributed to the literature on halo effects in student evaluations of teaching (SETs) by proposing and implementing a method to separate the effect of halo effects in student responses from an external measure of the item being assessed. Our paper has been criticised by Michela (2022). Many of his comments about problems with SETs are not directly relevant as they discuss issues other than halo. We re-visit our data and confirm that our conclusion that halo does not necessarily make SETs uninformative is correct. However, we do find heterogeneity in the importance of halo between SETs from two different campuses.

KEYWORDS

Student evaluation of teaching; validity; halo effects; lecture-room capacity

Introduction

In Cannon and Cipriani (2022), we published the results of an empirical analysis into halo effects in student evaluations of teaching (henceforth SETs). Since then, Michela (2022) has published a further paper in this journal consisting almost entirely of a critique of our paper. In this note we respond to Michela, whom we believe to have misunderstood the purpose of our paper. Our reading of Michela's paper is that we did not explain ourselves sufficiently clearly on some matters, and we are glad to re-state precisely what we are attempting to measure. We also take this opportunity to re-analyse the data in the light of Michela's comments and to provide further results. Our qualitative conclusions are unchanged. However, we find that the quality of the SET for the item that we are analysing varies between the two campuses for the university for which we have data.

The literature on SETs is huge. Since we reviewed the literature in our previous paper and since Michela (2022) also reviews the literature, we do not provide a literature review, but we note that the survey of Spooren, Brockx, and Mortelmans (2013) and the meta-analysis of Uttl, White, and Wong-Gonzalez (2017) both conclude that SETs can be unreliable measures of teaching ability.

When analysing SETs, it is important to distinguish two dimensions which are conceptually distinct, even if they may be related empirically. SETs might be used to measure or compare different tutors, i.e. to analyse *between*-teacher variation (to use SETs summatively to measure performance). Controversially, SET scores could be used in hiring, promotion or firing decisions: Becker, Bosshardt, and Watts (2012) provide evidence that SETs are used in this way in the U.S.A. SETs might also be used to measure or compare different aspects of teaching by the same tutor, i.e. to analyse *within*-teacher variation (to use SETs formatively or diagnostically to improve performance). To avoid any confusion or ambiguity, we emphasise that both our original paper and

CONTACT Edmund Cannon  edmund.cannon@bristol.ac.uk

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

this comment are entirely concerned with within-teacher variation. Purely for the purpose of record, we note that neither author has confidence in using SETs for between-tutor comparisons.

The purpose of our original paper

Nearly all SETs ask questions about a variety of aspects of a tutor's teaching. It is commonly observed that if an individual student gives a high score on one item, then the same student will tend to give a high score on other items and this correlation sometimes appears suspiciously high. This phenomenon is called the 'halo effect' and might suggest that a student's answer to one question is contaminated by answers to previous questions.

It is at least possible that students' responses to different items on the SET could form one type of feedback assisting a tutor to learn about their strengths and weaknesses as a tutor. Since there is no automatic connection between *within*-teacher variation in SET responses and *between*-teacher variation in SET responses, it remains possible that SET responses could assist tutors in improving their pedagogical practice, even if between-teacher variation in SET responses were completely invalid for comparing different tutors.

While there is no logical connection between between-teacher and within-teacher variation, it is possible that greater variation on one dimension could influence variation on the other. Our reading of Murphy, Jako, and Anhalt (1993) is that they think a stronger halo effect could magnify between-teacher variation and in our literature review we wrote: 'Although the halo effect reduces the reliability of within-teacher distinctions... by flattening the overall profile of ratings, on the other hand it can magnify differences in the mean ratings received by different teachers... It follows that the bias from a halo effect is not problematic if the purpose of SET is to distinguish a good teacher from a bad one, whilst it would be problematic if its purpose was to distinguish between strengths and weakness within a single teacher' (Cannon and Cipriani 2022, 3–4). Michela (2022, 3) quotes this text verbatim and interprets it to mean that we think that SETs can be used for between-teacher comparisons. We are sorry that it was unclear that this was merely our summary of Murphy, Jako, and Anhalt (1993) and we are happy to have the opportunity to clarify this point and to emphasise that this is not our own point of view.

Suppose we set aside completely the issue of between-teacher variation; if there is a large halo effect, then there will be minimal within-teacher variation, in which case the SET will be uninformative as a means for an individual tutor to identify their strengths and weaknesses.

Our original paper was written in the spirit of looking for reliability in within-teacher variation while remaining agnostic about the issue of between-teacher variation, and hence concentrated on the halo effect. As we have already stated, neither author has confidence in using SETs for between-teacher comparisons.

We now turn to the issue of measuring halo effects. An extreme possibility is that students give exactly the same answer to every question ('block rating'). It is implausible that teachers have identical ability on every aspect of teaching and so this is strong evidence for a halo effect. A simple analysis of our data shows that very few students engage in block rating, but correlations are very high. The purpose of our original paper – as suggested by the title – was to attempt to find some way of seeing how much of the positive correlation between responses was due to halo and how much was due to students genuinely responding to variation in the item being assessed.

To quantify the halo effect, one needs an independent measure of the item being assessed to compare with the student responses and in most cases no such independent measure exists. Notice that one could not use outcome measures such as student grades as this will depend upon all of the teaching resources provided in the unit and our objective is to analyse individual aspects of teaching; but, since SETs seem to have no relation to the quality of teaching, this issue is anyway irrelevant.

In our study we used data from the Italian university system. There are relatively few published studies of SETs in Italy, and the only examples that we found were Braga, Paccagnella, and Pellizzari (2014), Guerra, Bassi, and Dias (2020), Lalla, Facchinetti, and Mastroleo (2004), Lalla and Ferrari (2011) and Vanacore and Pellegrino (2019). None of these papers discussed halo effects. Since then, a further paper in this journal has shown that responses to different questions are correlated with each other, but did not draw any links to the issue of halo effects (Pastore, Manuti, and Scardigno 2022).

The Italian system requires students to respond on a Likert-scale of 1 to 4, giving fewer options than in the U.K. (five options) and many universities in the U.S.A. (which often use six or more). Since the possible responses are less granular in the Italian system this might have consequences for the degree of correlation between answers to different questions.

In our study we took advantage that the Italian system of SETs includes the question: 'Are the lecture theatres where this course is held adequate? Namely, can students see, hear, find a seat?' Teaching spaces are allocated so that the number of seats is at least as large as the number of students enrolled, but where 100 students are assigned to a room with exactly 100 seats it is very difficult for all students to find a convenient place to sit and the quality of the teaching session can vary. Furthermore, the quality of rooms can vary for other reasons (such as layout, quality of seats, quality of whiteboards or technical equipment and the acoustics of the room).

In the university for which we have data, rooms were allocated approximately randomly and teaching staff had no ability to ask to move room. This means that the quality of teaching space for a given unit should be uncorrelated with any aspect of the teaching due to the tutor. To the extent that more popular or better teachers might expect higher attendance and more over-crowding in lectures, there might even be a negative correlation.

Our initial discussion of halo is provided in Cannon and Cipriani (2022, Table 4). In our analysis we found strong positive correlations (between 0.26 and 0.34) between the answer to the room-quality question and the answers to other questions. Since there is no reason to expect a positive correlation, we are quite clear that this is strong evidence for a halo effect. Correlations between responses to the other questions were typically higher, lying in the range 0.27 to 0.68. Since teachers had more control over the other aspects of teaching, this is consistent with a positive correlation between ability on different dimensions of teaching but it is also consistent with a halo effect. The issue is whether there is any way to separate out the two explanations. In fact there is suggestive evidence that the responses to the questions are informative about something other than halo, even if what that 'something' represents is unclear: the Cronbach's alpha for different units is typically very high (Cannon and Cipriani 2022, 7). This means that if one student scores question 1 more highly than question 2 then other students are likely to do so too.

Many analyses in this field use a factor analysis. To make it easier to compare our paper to other pieces of research, we performed a factor analysis, which we viewed as a piece of exploratory data analysis (only devoting two paragraphs of text to the associated discussion). The factor analysis showed that the first factor had very high explanatory power. Using the commonly-used measure of the proportion of variation explained by the first factor, we found that it explained 97 per cent. This is further evidence that the answers to the different questions are jointly highly correlated. We noted that yet again the variables were all highly correlated, but that the room-quality question seemed to have a slightly lower correlation. We did not extend the factor analysis because we did not think that it added any additional insight to our results from the analysis of bivariate correlations.

We then moved to our original contribution, which was to use independent measures of room size (relative to students enrolled) and a dummy variable for students studying on different campuses (since the teaching rooms are quite different in the two campuses). In regression analysis we found that the proportions of variation explained by the information on room and campus were about twice as large as the proportion of variation explained by the responses to other questions.

From these results we concluded that the responses to this particular question were informative about the item being evaluated, despite being contaminated by halo effects. This suggests

that halo effects do not necessarily mean that SETs are completely uninformative when used for diagnostic purposes. Obviously, we do not know to what extent this is true of other questions or SETs in other universities. In an ideal world we should want independent measures of the other aspects of teaching to give more idea about the external validity of our results. But we did not claim universal validity, merely observing that SETs might continue to be informative for diagnostic purposes despite the issue of halo.

Econometric issues

Michela (2022) disagrees with our use of the campus dummy to explain variation in the students' responses to the question about rooms. In particular on page 9, he argues that this should not affect student responses because it is not a measure of room size. We accept that we highlighted the room-size issue and gave the title 'Identifying halo effects using room size' to our last section. With hindsight we explained this poorly and we thank Michela for drawing attention to this. However, it was always clear in our paper that the question asked of students was about the overall quality of the room. Two rooms might be of the same size and yet be of different quality in terms of visibility and audibility (explicitly mentioned in the question on the SET). Both co-authors have taught on both campuses and our personal experience led us to think that we should include this information.

Clearly there is scope to disagree *ex ante* on whether the campus variable should explain variation in student responses. Things become more difficult when one finds that the campus variable is highly statistically significant, even after conditioning on students' responses to other questions and the measure of room size. Before omitting this variable it would be necessary to explain why it is statistically significant and why it would be valid to exclude it from the analysis. The meaning of regression results that include the campus variable despite it being an invalid regressor are unclear, which makes the back-of-envelope calculations reported in Michela (2022, Table 2) difficult to evaluate. To complicate matters further, Michela's calculations do not take account that we use heteroskedasticity-consistent standard errors and F-statistics, so the numbers in his Table 2 are only approximately correct.

To cut through this Gordian knot, we report new regressions in Table 1. As a benchmark, the first specification is taken from Table 6 in our original paper (where it was specification 3). To avoid any debate over whether one should use the conventional or adjusted R-squared we report both, as they result in similar conclusions. Both the room capacity variable and the campus variable are highly significant and the unadjusted R-squared is 0.676; omitting the campus variable means that the R-squared falls dramatically to 0.456. This large fall in R-squared suggests that most of the work is being done by the campus variable, which might still make sense if the teaching spaces were better on one campus than the other. If one prefers the adjusted R-squared statistic, the results are qualitatively similar. The problem now is that we are pooling two groups of students being taught in different rooms of different qualities.

Table 1. Regression analysis for responses to question 13.

	(1)	(2)	(3)	(4)	(5)	(6)
Room capacity	0.198*** (0.041)	0.247*** (0.041)		0.336*** (0.077)		-0.015 (0.032)
Second campus	0.581*** (0.093)					
Responses to questions 1-12	Y	Y	Y	Y	Y	Y
N	61	61	37	37	24	24
R ²	0.676	0.456	0.422	0.650	0.435	0.440
adj R ²	0.578	0.306	0.133	0.452	-0.182	-0.287

Heteroskedasticity-consistent robust standard errors in parentheses (note that conventional standard errors give quantitatively similar conclusions). Specification (1) in this table is taken from Cannon and Cipriani (2022, Table 6, specification 3). * 0.05 ** 0.01 *** 0.001

Specifications 3 to 6 report results when we estimate regressions separately for each campus. It turns out that the SETs show very different behaviour in the two campuses. In the larger campus, the room-size variable remains significant (both statistically but also in terms of the magnitude of the parameter estimate): adding it to the regression raises the R-squared from 0.42 to 0.65. This is very strong evidence for our original conclusion: despite a halo effect, SETs correlate with an independent measure of the item being assessed. In the smaller campus, the room-size variable has no explanatory power (the parameter estimate is effectively zero). Given the R-squareds in regressions 3 and 5, the halo effect seems to be the same in both campuses. Again, using adjusted R-squareds gives qualitatively similar results.

It is a puzzle why there is a correlation in one campus but not the other. Given the size of the standard errors in both regressions, it is unappealing to appeal to Type I or Type II errors, small sample sizes notwithstanding: there appears to be a genuine difference between the two campuses.

Conclusion

We have responded to Michela's (2022) critique of Cannon and Cipriani (2022). We are grateful for the opportunity to clarify any ambiguity about precisely what the original paper was intending to demonstrate.

We have provided more detail for the argument in our original strategy of using both room size and a campus dummy to obtain an independent measure of room quality. In further analysis we have shown that the relationship between room size and the relevant SET response differs across the two campuses. This may be because the room size variable is a better proxy for quality in one campus than the other.

Our original conclusion was that the response to the room-quality question was contaminated by a halo effect but that it also provided a meaningful signal correlated with the independent measure of room quality. Hence halo effects are present but do not totally invalidate the use of SETs as a measure of within-teacher variation. In our new analysis here, we confirm that halo effects are present and that SET responses can be correlated with independent information.

Ethical considerations and declaration of interest

The data used in this study were obtained from an administrative database and anonymised by the university administrators who were not otherwise involved in the data analysis. It was not possible to obtain consent from the respondents and requesting informed consent might have influenced their answers: this issue was considered by the university ethics committee who gave special permission to use the data for this study, item 15908, dated 9 March 2015.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Neither author has any interest to declare.

Acknowledgements

We should like to thank the university administrators who provided data to us in an anonymised form. Any remaining errors are the authors' own.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Edmund Cannon is Professor of Economics at the University of Bristol.

Giam Pietro Cipriani is Professor of Economics at the University of Verona.

ORCID

Edmund Cannon  <http://orcid.org/0000-0002-1947-8499>

Giam Pietro Cipriani  <http://orcid.org/0000-0001-5436-0835>

References

- Becker, W. E., W. Bosshardt, and M. Watts. 2012. "How Departments of Economics Evaluate Teaching." *Journal of Economic Education* 43 (3): 325–333.
- Braga, M., M. Paccagnella, and M. Pellizzari. 2014. "Evaluating Students' Evaluations of Professors." *Economics of Education Review* 41: 71–88.
- Cannon, E., and G. P. Cipriani. 2022. "Quantifying Halo Effects in Students' Evaluation of Teaching." *Assessment & Evaluation in Higher Education* 47 (1): 1–14.
- Guerra, M., F. B., and J. G. Dias. 2020. "A Multiple-Indicator Latent Growth Mixture Model to Track Courses with Low-Quality Teaching." *Social Indicators Research* 147: 361–381.
- Lalla, M., G. Facchinetti, and G. Mastroleo. 2004. "Ordinal Scales and Fuzzy Set Systems to Measure Agreement: An Application to the Evaluation of Teaching Activity." *Quality and Quantity* 38 (5): 577–601.
- Lalla, M., and D. Ferrari. 2011. "Web-Based versus Paper-Based Data Collection for the Evaluation of Teaching Activity: Empirical Evidence from a Case Study." *Assessment & Evaluation in Higher Education* 36 (3): 347–365.
- Michela, J. L. 2022. "Toward Understanding and Quantifying Halo in Students' Evaluation of Teaching." *Assessment & Evaluation in Higher Education*.
- Murphy, K. R., R. A. Jako, and R. L. Anhalt. 1993. "Nature and Consequences of Halo Error: A Critical Analysis." *Journal of Applied Psychology* 78 (2): 218–225.
- Pastore, S., A. Manuti, and A. F. Scardigno. 2022. "A National Student Survey for the Italian Higher Education System." *Assessment & Evaluation in Higher Education* 47 (7): 985–997.
- Spooren, P., B. Brockx, and D. Mortelmans. 2013. "On the Validity of Student Evaluation of Teaching: The State of the Art." *Review of Educational Research* 83 (4): 598–642.
- Uttl, B., C. A. White, and D. Wong-Gonzalez. 2017. "Meta-Analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related." *Studies in Educational Evaluation* 54: 22–42.
- Vanacore, A., and M. S. Pellegrino. 2019. "How Reliable Are Students' Evaluations of Teaching (SETs)? A Study to Test Student's Reproducibility and Repeatability." *Social Indicators Research* 146: 77–89.