Routledge
Taylor & Francis Group

ASSESSMENT & EVALUATION in HIGHER EDUCATION
Volume 47 Number 4 June 2022

Assessment & Evaluation in Higher Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Toward understanding and quantifying halo in students' evaluation of teaching

John L. Michela

To cite this article: John L. Michela (2022): Toward understanding and quantifying halo in students' evaluation of teaching, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2022.2086965

To link to this article: https://doi.org/10.1080/02602938.2022.2086965

Published online: 20 Jun 2022.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2022.2086965

Toward understanding and quantifying halo in students' evaluation of teaching

John L. Michela

Department of Psychology, University of Waterloo, Waterloo, Canada

ABSTRACT

Despite its name, the halo effect in student evaluation of teaching (SET) response is not mystical. Halo in SET results from psychological processes that undermine SET validity, particularly for summative evaluation (pay decisions, etc.). These processes span psychological concepts of cognition, motivation and affect. This paper demonstrates that halo signals serious trouble for SET data, contrary to the conclusion in Cannon and Cipriani (2022), to which this article replies. Re-analysis of those authors' findings, conducted for the present paper, reveals both more halo and less validity than Cannon and Cipriani reported. For re-analysis of their factor analysis, a practical, generalizable approach to gauging halo in various settings is proposed. This approach revealed consistently high apparent SET halo across various decades and geographical locations of SET use. For re-analysis of their multiple regression, an approach to deciphering individual variables' effect sizes is provided for designs like Cannon and Cipriani's. Implications discussed for summative use of SETs include issues of veracity of SET confidence intervals. In contrast to summative use, formative use (toward performance improvement) suffers less from halo because non-instructional factor influences are somewhat constant for individuals.

KEYWORDS

Student evaluation of teaching; halo effect; cognitive schemas; effect size

Introduction

This paper provides a reply to the paper by Cannon and Cipriani's (2022), 'Quantifying halo effects in students' evaluation of teaching'. The present paper offers alternative interpretations of their findings, based partly on reanalysis of the statistical results presented in that paper, and partly on logical considerations. Contrary to their conclusion that 'distortion in the evaluation questionnaires caused by halo effects need not be a concern for higher education institutions' (1), the present paper draws implications from the nature and extent of halo effects which point to serious concerns for use of students' evaluations as measures of instructional performance.

The present paper is only partly in reply to Cannon and Cipriani (2022), in as much as it delves into the bases of students' survey responses in a fundamental manner. Halo can only be understood in relation to its associated psychological processes. The present paper also proposes novel statistical approaches to gauging halo, which may be applicable to multiple literatures.

The present paper begins with a basic description of the halo effect in survey response, within a section that provides an overview of what the focal paper does and does not show.

CONTACT John L. Michela jmichela@uwaterloo.ca

Then some of the bases of halo are sketched, expanding considerably beyond the focal paper's primary conception of the manifestation of halo as careless, block responding. Next, re-analysis of the focal paper's factor analysis indicates that halo effects were extensive, contrary to that paper's thrust. A further re-analysis, of the multiple regression concerning impacts of objective circumstances on SET ratings, uses algebraic transformations of conventional equations of regression analysis to show lower impacts upon students' ratings than claimed (as opposed to sizable impacts from perceptual distortion). We close with descriptions of how halo spuriously inflates the apparent accuracy of SET ratings.

What Cannon and Cipriani's study does and does not demonstrate

By definition, a halo effect is present when survey respondents' answers (or other ratings) are consistent with one another to an unwarranted extent. SET halo effects manifest as unrealistically high correlations among survey items.

Survey items may be correlated with one another either because of halo as a response tendency - a tendency that is a characteristic of the rater and thus unwarranted - or because of actual co-occurrence of features rated on the SET survey. The psychological literature often calls the former reason 'illusory halo' and the latter 'true halo'. 'Halo' means illusory halo in the present paper unless stated otherwise.

Halo's glow can extend quite far. Serra and McNeely (2020) posed, along with conventional SET questions, 'ridiculous questions'. These included: how likely is it that your instructor has a pet tiger; and how likely is it that your instructor's tears have magical healing powers? A composite score for these and other such items was calculated and then correlated with an overall rating of the instructor. The result was a substantial Pearson correlation (*r*) of 0.23.

Although Cannon and Cipriani acknowledged contamination of SET responses by halo effects in the data that they analysed, they concluded, from multiple regression analyses, that SETs are sufficiently valid for use in summative assessment. These analyses tested whether students' answers on a SET survey for an item about the classroom environment were associated with verifiable circumstances of the classroom environment. Figure 1, adapted from Fisicaro and Lance (1990), depicts the underlying conception for this analysis.

The survey question, Q13 in the figure, asked 'Are the lecture theatres where this course is held adequate? Namely, can students see, hear, find a seat?' The variables for the associated objective circumstances appear as Env1 and Env2 in the figure. Specifically, the authors' room capacity predictor variable, Env1, is essentially a ratio of classroom seats to enrolled students. Env2, called second campus in the focal paper, designates the campus location for each of the course offerings analysed here (first or second campus). Across a series of regression analyses, the authors analyse these two predictor variables individually and as a set, both with and without considering halo effects simultaneously.

The key finding was that room capacity, tailored by the researchers to constitute an external, objective representation of the classroom capacity as an environmental feature, was indeed

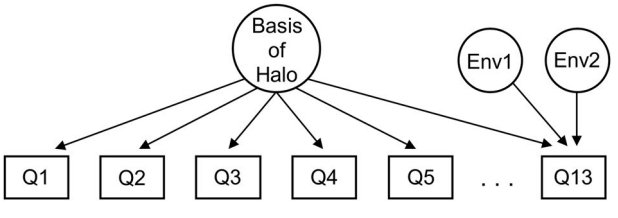

Figure 1. Model of halo effects in combination with effects of environmental circumstances or events for a designated survey item.

associated with responses to question 13. Furthermore, both halo and the two environmental predictors jointly contributed to the prediction of responses to question 13.

The authors are warranted in concluding that these findings provide some external validation for responses to question 13. Further, the authors properly concur with a comprehensive review and analysis of halo from Murphy, Jako, and Anhalt (1993) in holding that presence of halo does not necessarily invalidate survey responses.

However, apart from technical/statistical issues that will be covered in a later section, there are immediate conceptual or logical issues for the authors' conclusion about overall SET validity on the basis of their findings. The survey item at the heart of the corresponding empirical finding, is *least* capable of sustaining such a broad conclusion. To repeat, question 13 asks whether students can 'see, hear, find a seat'. Cooper's (1981, 220) widely cited review of halo effects held that we should expect least halo error for survey items like this one, involving matters that are more 'concrete' versus abstract. Cooper classifies rating categories as concrete when they are 'highly descriptive, empirically derived, and sufficiently specific and concrete, as opposed to abstract'.

Similarly, Feeley's (2002, 582) review cites 'explicitness' as a factor in halo:

Explicit attribute types use counts and amounts and leave little room for rater interpretation, thereby reducing halo error. Inferential attribute types [in contrast] require raters to make judgments about the meaning and relevance of target behaviors.

It follows that the authors' findings for the concrete/explicit matter of classroom setup tells us *little about rating dimensions concerning instruction*, which involve survey items that pose abstract and inferential questions. Rater judgments or evaluations are inherently abstract and inferential when they concern matters such as 'adequate' teaching and learning materials to study the subject (as mentioned in the authors' item 5), 'clear' information about examination structure (item 6), 'adherence' to teaching activities (item 9), and 'clear' explanations (item 11).

The question of how the authors could have offered their broad dismissal of trouble from halo on this limited basis might be illuminated by examining their statement:

Although the halo effect reduces the reliability of within-teacher distinctions [in formative evaluation use of SETs] by flattening the overall profile of ratings, on the other hand it can magnify differences in the mean ratings received by different teachers [for use in summative evaluation]. 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).

Here the authors have erroneously 'assumed the conclusion' (Wikipedia 2021) that halo in SETs does not impair their validity for summative use. If halo were merely to involve imprecision in differentiating one's *otherwise-valid answers*, then the authors' conclusion might hold up. Indeed, Murphy, Jako, and Anhalt (1993) allow this possibility of score enhancement from halo when supervisors rate employees' work performance, because supervisors have extensive knowledge about the job and worker.

But the context of students' evaluations of teaching is fundamentally different. Students do *not* base their ratings on how effectively they have been taught, according to literature reviews by Boring, Ottoboni, and Stark (2016), Stroebe (2020), and Uttl, White, and Gonzalez (2017). Other cognitive, motivational and affective bases of students' ratings *have* been demonstrated. These bases are tied to non-instructional factors that are as diverse as the weather (e.g. Braga, Paccagnella, and Pellizzari 2014) and students' retaliation for low marks (e.g. Clayson and Haley 2011). These bases produce halo effects.

What are the bases of halo effects?

By definition, a halo 'effect' is an outcome of some process or processes. Because SET survey data are provided by people, these processes are psychological in nature. The part of the causal

model of Figure 1 that traces back to a 'basis' of halo is a place holder for any number of such processes. Intercorrelations are unwarrantedly high among survey items, and therefore indicative of illusory halo, when answers are not as differentiated as they would be if they were more tied to environmental realities (such as classroom conditions, instructor behaviour, etc.).

Coverage of bases or sources of halo (e.g. Cooper 1981) usually focus on form over content of halo's basis. The present coverage is novel, as it will draw on social and industrial-organizational psychology to describe specifics of content in psychological processes that increase halo in the SET context.

Cognitive schema-based halo

Cognitive schemas are mental structures that guide perception and other mental processing in some defined stimulus domain (Taylor and Crocker 1981). Perception and evaluation in work contexts is one such domain in which schemas are understood to operate (e.g. DeNisi, Cafferty, and Meglino 1984). For the performance evaluation context, schemas encode knowledge or beliefs about aspects or attributes of high versus low performance. For example, one's schema for high performance by a comedian may include 'makes me laugh' and 'makes witty observations about human nature'. Taylor and Crocker add that a schema includes specification of the relationships among its attributes. This matter of what-goes-with-what is central to halo.

This schema theory indicates that students' schemas can influence what is encoded into memory, and therefore possible at all to recall, as well as how recalled information is integrated toward SET judgment. And when information is scarce or ambiguous for some survey items (e.g. lacking concreteness), as it often is, schema theory indicates that the schema provides direction for how to fill in the gaps to be consistent with a stereotype, prototype or exemplar. DeNisi, Cafferty, and Meglino (1984, 369) provide an extensive model of information processing in performance appraisal which illuminates the gap filling process:

A rater who has categorized a worker in terms of a 'good worker' schema, based on a few salient observations may feel he or she 'already knows' how good a job the worker is doing and not see the need to collect much additional information about that worker. Furthermore, what additional information is sought may be distorted to conform with the rater's preconceived notions.

Thus, halo is magnified when schematic processing of performance information works in concert with general cognitive biases and heuristics such as anchoring, availability and confirmation bias (Parks 2018).

Stereotype of communal women and agentic men

SET rating bias that favours male over female instructors is most often understood to have its basis in the traditional schematic gender stereotypes (e.g. Basow, Phelan, and Capotosto 2006). Experimental evidence for the warmth-competence stereotype as a basis for SET rating differences comes from Graves, Hoshino-Browne, and Lui (2017). An identical physics lecture was presented either by a male or female actor. Both male and female students evaluated male professors' scientific knowledge and skills more highly. Female students evaluated female professors' interpersonal skills more positively.

Stereotypes are a form of schema (Taylor and Crocker 1981). Therefore, the gap filling process described earlier (DeNisi, Cafferty, and Meglino 1984) can be expected to produce halo when men's ratings become inflated (or women's are decremented) based on assumed competence. Women who violate interpersonal warmth can pay a SET price (Sinclair and Kunda 2000).

Status incongruity with group stereotypes

Status incongruity exists when a person from a lower status group occupies a higher status position. Fisher, Stinson, and Kalajdzic (2019, 305) explain that 'this perceived incongruity may

threaten people's belief that the world operates according to a set of just and fair rules, rules that include gender-role expectancies'. In this theory, negative evaluations seek to re-establish a status quo of lower status for women. Consequently, reduced differentiation among SET ratings on various dimensions can be expected, thus magnifying halo.

Miller and Chamberlin (2000) document lower status perceptions for female instructors, in terms of assumptions about relative rates of holding a Ph.D. degree across genders. Further, occurrences of racial or cultural SET bias may stem from status incongruence. The magnitude of this bias may generally exceed gender bias (e.g. Smith and Hawkins 2011), though evidence is limited for making this comparison.

Schemas for 'good instruction'

Other schemas tell students what good *instruction* looks like. 'Sage on the stage' is a self-explanatory label for a salient schema (e.g. McWilliam 2008) that involves attributes such as confident delivery of lecture material and enthusiasm. When this schema operates, schema theory predicts that SET item responses will shift into conformance with one another. For example, if a student thinks the instruction has been especially 'good' on any basis, she will tend to rate the instructor's explanations as truly 'clear', the teaching and learning materials as 'adequate' and so forth (thus producing halo).

Unfortunately, 'sage' instruction is counter to often-recommended instructional practices such as active learning. In Deslauriers et al. (2019) students were randomly assigned to traditional or active learning instruction. Students in active learning classes provided lower SET ratings even though those students learned the most, based on test scores. Moreover, ratings of *perceived learning* by students in the active learning group were *lower* than in the control group.

Findings from Deslauriers et al. are consistent with several differently designed studies, reviewed by Stroebe (2020). These studies examined students' course ratings along with performance on examinations across structured course sequences, such as Calculus I and Calculus II. In five of six such studies, students who earlier had been in higher rated courses turned out to obtain lower marks in the later course. As Braga, Paccagnella, and Pellizzari (2014, 81) summarize, instructors 'who are more effective in promoting future performance receive worse evaluations from their students'.

Motivated responding-based halo

Some students consciously use SET ratings to reward or punish instructors regarding either course workload or marks (Stroebe 2020). Clayson and Haley (2011, 104) summarize a student survey in three universities which found:

> 30 percent admitted to purposely inflating evaluations beyond what was deserved because an instructor gave good grades; another 30 percent indicated that they had purposely lowered evaluations below that which was deserved because the tests in the course were 'too hard'. Fifty percent had done one or the other. Eighty percent… indicated that they had knowingly given an instructor an undeserved evaluation for some reason.

When this motivation is operating, the pattern of response across items will narrow to more uniformly high or low ratings (to accomplish the desired end), magnifying halo.

Affect-based halo

Many other factors that lead to halo evidently operate though students' feelings about course attributes and about the instructor. The process involves transferred affect from reactions to the situation (events, circumstances, other people) to the ratings provided. Effects of weather were mentioned earlier. Nilson (2012) lists many others (Asian accent, class length, classroom

pleasantness, etc.). Zumbach and Funke (2014) provide direct evidence of impact from affect. They observed more favourable overall ratings of quality of the course among students in an induced positive mood condition compared with a negative mood condition.

Instructor likability is especially potent. In Murray, Rushton and Paunonen's study (1990), 29 instructors of introductory psychology were rated by their departmental peers in terms of various personality traits. Across these 29 instructors, correlations were computed between average faculty *peer* ratings of sociability and average student ratings of the course, resulting in a considerable $r = 0.64$. Extraversion yielded the same result.

As with schemas, affect can be expected to magnify halo. Ratings of attributes that are least tied to the source of the affect will be brought into conformance when external referents are few and inference is high.

Quantifying halo in the focal paper's complete set of survey items

Although Cannon and Cipriani (2022, 8) report 'little evidence of extreme halo effects since the prevalence of block grading is low', conspicuously high halo remains.

Bivariate associations

The focal paper's Table 4 provides bivariate associations among all SET survey items. These correlations are somewhat low relative to other empirical articles. The median correlation in Table 4 is 0.38. In Johnson, Narayanan, and Sawaya (2013) the median inter-item correlation was 0.80. Williams and Ceci (1997) report intercorrelations separately among instructor-related items (mean $r = 0.53$) and course-related items (mean $r = 0.56$). The present author's estimate is that the median of just below 0.60 that was reported at his university (WCPS Team 2020) is typical. The university chose its nine survey items to have some degree of commonality, to achieve reliable measurement of constructs, along with some degree of differentiation, for content validity.

Because different instruments take different approaches to SET item selection, and obtain different levels of inter-item correlation, a generalizable multivariate gauge of halo is needed.

Multivariate associations

The authors' factor analysis in Table 5 sought this multivariate quantification. The authors highlight how the first of five extracted factors explained 97% of the variance in the associations among survey items. Although there is precedent for attempting to quantify halo this way, it does not withstand scrutiny (Murphy, Jako, and Anhalt 1993). Further, Comrey and Lee (2013) state that multiple factors that underlie survey responses are almost always obscured in an unrotated factor solution.

Accordingly, the focal paper's five unrotated factors were analysed by the present author to obtain non-orthogonally rotated factors (i.e. correlated factors). Specifically, the factor structure in the authors' Table 5 was entered as matrix data into IBM SPSS Statistics for Windows (version 21.0), requesting promax rotation for 5, 4 and 3 factors. The present paper's Table 1 provides the resulting three-factor rotation of the focal paper's factors (rotated factors four and five were barren, with no survey questions having loadings above the conventional cut-off of 0.40).

The composition of Factor 1 supports the authors' stated view that the first four items are mostly distinct from the rest, involving matters mostly external to the course. All four items loaded there. Accordingly, at the bottom right of Table 1 this factor is labelled 'Program context for course'. Factor 2 has been labelled 'Operational matters' as this factor captures examination

Table 1. Factor loadings (promax pattern matrix) and correlations based on first three factors of Table 5 in Cannon and Cipriani (2022).

Item Number	F1	F2	F3	Item Content
Q1	***0.80***	−0.04	−0.02	Overall study load
Q2	***0.55***	0.24	−0.04	Organisation
Q3	***0.45***	−0.09	0.24	Adequacy of prior knowledge
Q4	***0.63***	0.07	0.04	Study load and credits
Q5	0.26	0.29	0.25	Teaching materials
Q6	0.03	***0.61***	0.09	Examination information
Q7	−0.04	***0.65***	0.14	Teacher availability
Q8	0.17	−0.05	***0.55***	Topic interesting
Q9	−0.02	***0.74***	−0.03	Adherence to timetable
Q10	−0.02	0.06	***0.79***	Teacher motivates students
Q11	−0.03	0.11	***0.76***	Teacher's explanation
Q12	0.02	***0.66***	0.14	Consistency with website
Q13	0.18	***0.41***	−0.10	Lecture room
Factor Intercorrelations				*Factor Content*
F1	1.00			Program context for course
F2	0.71	1.00		Operational matters
F3	0.68	0.80	1.00	Cognitive-motivational
	F1	*F2*	*F3*	

information, instructor availability and consistency of course components. Not surprisingly, question 13 on room characteristics also appears here. Factor 3 is labelled 'Cognitive-motivational impacts' as it involves interest, motivation and explanation clarity.

The intercorrelations of these *factors* provides a superior basis for gauging halo. These correlations, as shown at the bottom left of Table 1, are very sizable. A necessary caveat for gauging halo this way is that the relative extent of true versus illusory basis of associations for factors remains indeterminate, just as with survey *items*. Interpretive judgment is required. For the focal study, this interpretation requires consideration of the plausibility of true halo as the source of correlations; of magnitudes of inter-factor correlations versus other reference points; and of the concrete versus inferential nature of survey items.

Given the authors' statement that the first four items are mostly distinct from the rest, there is little basis for true halo to account for the correlations of Factor 1 with Factors 2 and 3. The corresponding correlation values that are therefore attributable mostly to illusory halo are high, at 0.71 and 0.68. As a point of reference, these values' corresponding proportions of shared variance, 0.50 and 0.46 (i.e. the correlation values squared), may be compared with the authors' reported amount of variance in survey response attributable to halo. This amount is approximately 0.20 according to the subsequent multiple regression analyses in the focal paper. Thus, this figure of 0.20 appears to be a considerable understatement of the extent of illusory halo as compared with 0.50 or 0.46, even considering that some portion of the latter figures may constitute true halo.

Use of factor analysis provides this divergent perspective partly because correlations among factors are between latent variables, which are effectively corrected for attenuation from measurement error and thus are higher than otherwise. In contrast, the focal paper's Table 4 provides correlations at the level of the survey items, inherently imbued with measurement error. The correlations in Table 4 imply less than 0.15 as the typical proportion of shared variance between survey items. Again, this figure far understates likely illusory halo for the more meaningful level of underlying constructs (these underlying constructs are described tersely by the labels for the factors). In emphasizing associations at the construct level, the present approach takes its cue from meta-analysis practice, which commonly employs correction for attenuation due to measurement error (Schmidt and Hunter 2014).

Notably highest in the present Table 1 is the correlation between the authors' Factors 2 and 3, at $r = 0.80$. Each of these factors involves predominantly interpretive matters, allowing for a great deal of illusory halo. Interpretation of the association, instead, as largely from true halo,

would require belief that courses with the favourable cognitive-motivational attributes or outcomes of Factor 3 are overwhelmingly ones that also have the favourable operational attributes of Factor 2. This is possible, though highly questionable.

Similarity with other such studies' findings

Although the authors' inter-item correlations were low relative to those of many other researchers, their inter-factor correlations were very similar to those from some published or public sources obtained by the present author. These comparative findings are provided in Michela (2022a), which serves as Supplement one to the present manuscript. The similarities across studies, including the focal study, are striking in terms of the inter-factor correlations, ranging rather narrowly between 0.67 to 0.80. Although only these few studies have been included, they span time (more than 40 years) and geographical locations (Canada, Italy, USA).

Some of the inter-factor correlations in these additional studies similarly appear to reflect considerable illusory halo. For example, in the first study in Supplement one (reported in Table A) a factor that mainly concerned communication and alignment of course components with learning objectives yielded a correlation of .73 with presumably more central evaluations of communication clarity, interest stimulation and perceived learning. It is difficult (though not impossible) to imagine that such communication and alignment with learning objectives is coincident to this extent with the other course attributes (assuming that learning objectives, per se, are not salient to students). In this same study it is also noteworthy that the most *concrete* item, concerning timing of return of assignments, is least associated with the other items (showing low loadings on both factors). Findings in the study of Table B include an implausibly high correlation (0.80) between course content preference and ratings of the instructor, although other interpretations are acknowledged in Supplement one.

Table C was included to illustrate how illusory halo may not dominate inter-factor correlations in some instances. Specifically, the high correlations between the factor for overall evaluations and the more specific factors may reflect a logical connection between overall evaluation and the specifics on which the overall assessment is based.

Quantifying halo relative to veridical perception

'Veridical' perception is that which is solidly tied to external reality. This section disputes the authors' claim that their multiple regression analyses show that identifiable, objective factors in students' classroom environment are dominant over halo in students' responses to survey question 13 about classroom adequacy, despite this survey item's concrete nature.

Close examination of these findings requires a common metric of effect size for the influences of objectively assessed classroom conditions versus halo. The authors used comparisons of overall R^2 values across the various regression models to provide this metric. However, it is also necessary to partition the explained variance (R^2) within each model.

The present author attempted this partitioning within the constraint of basing it on the regression analysis findings in the focal paper's Table 6. This undertaking required various assumptions and extrapolations. First, the regression coefficients are assumed unstandardized, which is the form that most requires further analysis. If instead they are standardized, the overall upshot of this section is very likely to stand, because the coefficients for room capacity would be the smaller of the two when both coefficients are in the standardized metric. Second, the reported R^2 values are assumed to have been adjusted R^2, given that statistical significance otherwise would not have been reached for some of the reported R^2 values (focus on adjusted R^2 is warranted because the large number of predictors for halo allows shrinkage from R^2). Third, the present re-calculations applied the fully conventional regression model (e.g. Cohen and Cohen 1983), although (a) the authors mention their use of robust standard errors, and (b)

Table 2. Quantities of variance explained in survey question 13 by two individual predictors and by the two predictor blocks.

	Model 1		Model 2		Model 3		Model 4	
Objective influence estimators								
Block explained variance estimate as R^2_change	0.52	[1]			0.40	[1]	0.58	[1]
Corresponding adjusted R^2	0.50	[2]			0.38	[2]	0.56	[2]
F-ratio basis of estimates of explained variance	F (2, 58) = 30.80	[3]			F (2, 46) = 37.30	[3]	F (2, 50) = 49.30	[3]
Room capacity unique variance (sr^2)	0.08	[4]			0.13	[4]	0.16	[4]
Second campus unique variance (sr^2)	0.37	[4]			0.21	[4]	0.24	[4]
Halo influence estimators								
Block explained variance estimate as R^2_change			0.33	[1]	0.31	[1]	0.24	[1]
Corresponding adjusted R^2			0.17	[2]	0.13	[2]	0.13	[2]
F-ratio basis of estimates of explained variance			F (12, 48) = 2.00	[3]	F (12, 46) = 4.70	[3]	F (8, 50) = 5.20	[3]
Overall model statistics								
Overall R^2	0.52	[1]	0.33	[1]	0.75	[5]	0.71	[5]
Overall Adjusted R^2	0.50	[2]	0.17	[2]	0.68	[6]	0.65	[6]
Estimated Conventional Model Overall F-ratio					F (14, 46) = 9.93	[7]	F (10, 46) = 12.09	[7]

Note. For purposes of comparison with the statistically controlled, sr^2 values in this table, the simple Pearson correlation r and r^2 values, based on equations in Supplement two, are as follows: Room capacity $r = 0.38$; $r^2 = 0.14$. Second campus $r = 0.66$; $r^2 = 0.44$.

Legend

[1] R^2 that would result in conventional multiple regression (CMR) for the corresponding F-ratio in the focal paper's Table 6.

[2] Adjusted R^2 for the immediately preceding (above) CMR R^2 .

[3] F-ratio as provided in Table 6.

[4] sr^2 value calculated as described in Supplement two.

[5] R^2 corresponding to immediately following (below) adjusted R^2 value.

[6] Adjusted R^2 value in Table 6.

[7] Estimated F-ratio in CMR that would yield this model's corresponding overall adjusted R^2 as reported in Table 6.

some of the relations between reported elements (R^2, *F*-ratio, *df*) are not completely consistent with conventional regression analysis. While imprecise, the present application of the conventional multiple regression model allowed approximation of the needed quantification, by use of the equations that have been collected and algebraically transformed in Michela (2022b), a source that serves as Supplement two to the present paper. Details of the reanalysed findings appear in the present paper as Table 2, presented in an arrangement resembling the authors' Table 6.

The important further findings for the authors' Model 1 involve the *separate* contributions of room capacity and second campus. These variables require differentiation because the room capacity predictor variable, as described earlier, is most specifically pertinent to whether students can 'see, hear, find a seat?' in question 13. The authors had stated: 'The main contribution of this paper is to identify halo effects through the question on lecture-room size'. Yet all the cross-model comparisons involve room capacity and second campus *jointly*. The authors refer to the second campus variable as a *control* variable, as in their statement that 'some units were lectured at a different campus and we control for this with an indicator dummy variable C_u'.

Equation (5) in Supplement two provides a basis for determining effect sizes separately for room capacity and second campus, in the form of each predictor's sr^2 (i.e. semi-partial r^2). This parameter tells the outcome variable variance that is explained uniquely by a predictor. To begin the present re-analysis that is detailed in Table 2, each predictor's *t*-ratio was calculated from the regression coefficient and standard error for Model 1 in the focal paper's Table 6. The *t*-ratio was squared (following equation 6 in Supplement two) to generate the *F*-ratio for use in equation (5).

The resulting effect size estimate for room capacity is $sr^2 = 0.08$. This value is much smaller than the R^2 of 0.47 reported by the authors for room capacity and second campus jointly. Indeed, this effect size for room capacity is *smaller* than effect sizes for *halo* that emerge in the regression models. The focal paper attributes approximately 20% of variance to halo (e.g. $R^2 = 0.20$ in Model 2). The present reanalysis concurs in this regard, as re-analysis of Model 2 provides a sufficiently similar halo effect size estimate (i.e. adjusted $R^2 = 0.17$). As compared with re-analysed Model 1, reanalysis of Model 3 and Model 4 yields somewhat larger values for room capacity ($sr^2 = 0.13$ and 0.16). However, these values for the objective indicator remain less than the authors' initial figure of 0.20 for halo or the re-analysis figure of 0.17. Furthermore, in Model 3 and Model 4 halo yields adjusted R^2 values of 0.13.

Considering all these findings, room capacity *cannot* be said to explain more variance than halo. It is only with addition of the 'control' variable, second campus, that the *block* of the room variables is dominant over halo. Moreover, room capacity appears to be associated *less* strongly with question 13 than second campus, even though room capacity was tailored to align with question 13. It is difficult to picture how veridical perception adequately accounts for these findings.

Consistent with Nilson's (2012) review and analysis of non-instructional factors that impact students' evaluations of teaching, affective differences between the campuses may provide a better account. One campus may be more convenient, more enjoyable in terms of course types or amenities, and so forth.

Implications

Halo effects in SET ratings signal deficient validity. However, the nature and consequences of this deficiency differ between summative and formative uses of SET.

Implications for summative use of SET

There is a straightforward requirement for SET summative validity. It is that instructors who *perform* better as instructors must receive SET *ratings* that are higher than those of their lower performing colleagues. This is not what happens. Non-instructional factors that produce halo have the effect of scrambling up the order of instructors' SET ratings relative to actual instructional performance. That is, halo magnifies differences among instructors that are *not* tied to actual performance, by increasing the internal consistency reliability of instructors' scores on invalid bases.

Simultaneous with this reduction in validity, halo-induced increases in reliability produce distortion in the confidence intervals (CIs) that take account of measurement error (as in Benton and Li 2017). Higher reliability, warranted or not, reduces the standard error of measurement (SEM) which in turn makes CIs narrower. The root of the problem is that CIs describe the range of likely true scores for *whatever SET instruments measure*. The catch is that 'student evaluations of teaching do not measure teaching effectiveness' according to an up-to-date, comprehensive review of SET validity and bias (Stroebe 2020, 283). Disturbingly, when the class average SET score that an instructor receives is lacking in validity after the operation of non-instructional factors and halo, reporting of these SEM-based CIs is worse than useless; it is misleading.

Instead of SEM-based CIs, many SET administrators provide *sampling error*-based CIs. These CIs look better (i.e. narrower) with higher response rates, so administrators sometimes seek to overcome response reticence by use of lotteries or other inducements. It is reasonable to expect that students motivated in this way will minimize effort, as through doing block responding, or through applying stereotypes, relying on their feelings about the course, and so forth. Lower validity would result.

Implications for formative use of SET

Contrary to the focal paper's conclusions, halo and its bases are not as damaging to SET validity for formative use as they are for summative use. Three kinds of comparisons made during interpretation of SET scores are pertinent here.

First is the comparison of scores across people, as when department members are assigned relative performance scores. For formative use, this comparison has little or no pertinence when it is understood that non-instructional factors are the primary determinants of instructors' SET ratings and thus rankings.

Next there is comparison across SET survey items. A key purpose of using SET in formative evaluation is to identify areas or dimensions of performance to prioritize for performance improvement. Yes, halo is detrimental here, because any individual's variation from lower to higher SET scores is unwarrantedly restricted when there is halo. However, there is no reason to expect halo to produce scrambling of the relative standings of the scores in an instructor's profile, comparable to scrambling of instructors' standings relative to one another in summative evaluation. Instructors and coaches in teaching support services should be aware of this within-instructor restriction and try to take it into account when interpreting SET scores.

Finally, there is the comparison across time for a given instructor. SET scores can assess change from modifying instructional practice. Again, halo is detrimental but not devastating. If targeted intervention that is limited to few dimensions yields increases in many dimensions, useful information still has been obtained about intervention impact. Further, although correlations tend to be high between items or scales, they are not so high as to move dimensions in lockstep. A study by Williams and Ceci (1997) is relevant. Instructional behaviour change, as coached by a teaching support service, produced increases in all assessed dimensions in this study. However, the instructor's previously most deficient dimension - the target of the intervention - increased most noticeably.

Conclusion

Cannon and Cipriani (2022) are not alone in their optimistic assessment of halo. The author of the earliest cited work on halo, Wells (1907, 20), said the problem of halo was not 'serious'. But context matters a lot with performance ratings (DeNisi and Murphy 2017). From all we know about the likely magnitude of illusory halo, and about bases of SET response, the image to associate with SET is not that of a halo but, instead, that of a pitchfork, horns and cloven hooves.

Acknowledgements

The author sincerely thanks Ramona Bobocel, Beth Jewkes, and Erik Woody, both for insightful comments on an earlier draft of this manuscript and for continuing support of the author's related research.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

John (Jay) Michela recently transitioned to Professor Emeritus in the Department of Psychology, University of Waterloo. He co-founded the doctoral program there in Industrial-Organizational Psychology, and previously he held the position of tenured Associate Professor in Social and Organizational Psychology at Teachers College, Columbia University, after receiving his Ph.D. at the University of California, Los Angeles.

References

Basow, S. A., J. E. Phelan, and L. Capotosto. 2006. "Gender Patterns in College Students' Choices of Their Best and Worst Professors." *Psychology of Women Quarterly* 30 (1): 25–35. doi:10.1111/j.1471-6402.2006.00259.x.

Benton, S. L., and D. Li. 2017. *IDEA Student Ratings of Instruction and RSVP*. Idea paper #66. Manhattan, KS: The IDEA Center.

Boring, A., K. Ottoboni, and P. B. Stark. 2016. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." In *ScienceOpen Research*, 1–11. https://www.scienceopen.com/document/read?vid=818d8ec0-5908-47d8-86b4-5dc38f04b23e.

Braga, M., M. Paccagnella, M., and Pellizzari, M. 2014. "Evaluating Students' Evaluations of Professors." *Economics of Education Review* 41: 71–88. doi:10.1016/j.econedurev.2014.04.002.

Cannon, E., and G. P. Cipriani. 2022. "Quantifying Halo Effects in Students' Evaluation of Teaching." *Assessment & Evaluation in Higher Education* 47 (1): 1–14. doi:10.1080/02602938.2021.1888868.

Clayson, D. E., and D. A. Haley. 2011. "Are Students Telling us the Truth? A Critical Look at the Student Evaluation of Teaching." *Marketing Education Review* 21 (2): 101–112. doi:10.2753/MER1052-8008210201.

Cohen, J., and P. Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Comrey, A. L., and H. B. Lee. 2013. *A First Course in Factor Analysis*. Hove, UK: Psychology Press.

Cooper, W. H. 1981. "Ubiquitous Halo." *Psychological Bulletin* 90 (2): 218–244. doi:10.1037/0033-2909.90.2.218.

DeNisi, A. S., and K. R. Murphy. 2017. "Performance Appraisal and Performance Management: 100 Years of Progress?" *The Journal of Applied Psychology* 102 (3): 421–433. doi:10.1037/apl0000085.

DeNisi, A. S., T. P. Cafferty, and B. M. Meglino. 1984. "A Cognitive View of the Performance Appraisal Process: A Model and Research Propositions." *Organizational Behavior and Human Performance* 33 (3): 360–396. doi:10.1016/0030-5073(84)90029-1.

Deslauriers, L., L. S. McCarty, K. Miller, K. Callaghan, and G. Kestin. 2019. "Measuring Actual Learning versus Feeling of Learning in Response to Being Actively Engaged in the Classroom." *Proceedings of the National Academy of Sciences of the United States of America* 116 (39): 19251–19257. doi:10.1073/pnas.1821936116.

Feeley, T. H. 2002. "Comment on Halo Effects in Rating and Evaluation Research." *Human Communication Research* 28 (4): 578–586. doi:10.1111/j.1468-2958.2002.tb00825.x.

Fisher, A. N., D. A. Stinson, and A. Kalajdzic. 2019. "Unpacking Backlash: Individual and Contextual Moderators of Bias against Female Professors." *Basic and Applied Social Psychology* 41 (5): 305–325. doi:10.1080/01973533.2019.1652178.

Fisicaro, S. A., and C. E. Lance. 1990. "Implications of Three Causal Models for the Measurement of Halo Error." *Applied Psychological Measurement* 14 (4): 419–429. doi:10.1177/014662169001400407.

Graves, A. L., E. Hoshino-Browne, and K. P. Lui. 2017. "Swimming against the Tide: Gender Bias in the Physics Classroom." *Journal of Women and Minorities in Science and Engineering* 23 (1): 15–36. doi:10.1615/JWomenMinorScienEng.2017013584.

Johnson, M. D., A. Narayanan, and W. J. Sawaya. 2013. "Effects of Course and Instructor Characteristics on Student Evaluation of Teaching across a College of Engineering." *Journal of Engineering Education* 102 (2): 289–318. doi:10.1002/jee.20013.

McWilliam, E. 2008. "Unlearning How to Teach." *Innovations in Education and Teaching International* 45 (3): 263–269. doi:10.1080/14703290802176147.

Michela, J. L. 2022a. *Use of Non-Orthogonal Factor Analysis for Gauging Illusory Halo: A Technical Report*. Waterloo, Canada: UWSpace Document Repository, University of Waterloo. http://hdl.handle.net/10012/18340.

Michela, J. L. 2022b. *Equations for Deriving Effect Sizes for Individual Predictors and Sets of Predictors under Specified Conditions in Multiple Regression Analysis: A Technical Report*. Waterloo, Canada: UWSpace Document Repository, University of Waterloo. http://hdl.handle.net/10012/18341.

Miller, J., and M. Chamberlin. 2000. "Women Are Teachers, Men Are Professors: A Study of Student Perceptions." *Teaching Sociology* 28 (4): 283–298. doi:10.2307/1318580.

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. doi:10.1037/0021-9010.78.2.218.

Murray, H. G., J. P. Rushton, and S. V. Paunonen. 1990. "Teacher Personality Traits and Student Instructional Ratings in Six Types of University Courses." *Journal of Educational Psychology* 82 (2): 250–261. doi:10.1037/0022-0663.82.2.250.

Nilson, L. B. 2012. "Time to Raise Questions about Student Ratings." In *To Improve the Academy*, Vol. 31, edited by J. E. Groccia and L. Cruz Castro, 213–227. Hoboken, NJ: Wiley. doi:10.1002/j.2334-4822.2012.tb00683.x.

Parks, G. S. 2018. "Race, Cognitive Biases, and the Power of Law Student Teaching Evaluations." *University of California at Davis Law Review (UCDL Rev)* 51 (3): 1039–1079.

Schmidt, F. L., and J. E. Hunter. 2014. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 3rd ed. Thousand Oaks, CA: Sage.

Serra, M., and D. A. McNeely. 2020. "The Most Fluent Instructors Might Choreograph for Beyoncé or Secretly Be Batman: Commentary on Carpenter, Witherby, and Tauber." *Journal of Applied Research in Memory and Cognition* 9 (2): 175–180. doi:10.1016/j.jarmac.2020.02.005.

Sinclair, L., and Z. Kunda. 2000. "Motivated Stereotyping of Women: She's Fine If She Praised Me but Incompetent If She Criticized Me." *Personality and Social Psychology Bulletin* 26 (11): 1329–1342. doi:10.1177/0146167200263002.

Smith, B. P., and B. Hawkins. 2011. "Examining Student Evaluations of Black College Faculty: Does Race Matter?" *Journal of Negro Education* 80 (2): 149–162. https://www.jstor.org/stable/41341117.

Stroebe, W. 2020. "Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis." *Basic and Applied Social Psychology* 42 (4): 276–294. doi:10.1080/01973533.2020.1756817.

Taylor, S. E., and J. Crocker. 1981. "Schematic Bases of Social Information Processing." In *Social Cognition. Vol. 1 of the Ontario Symposium*, edited by E. T. Higgins, P. Herman, and M. Zanna, 89–134. Hillsdale, NJ: Lawrence Erlbaum.

Uttl, B., C. A. White, and D. W. 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. doi:10.1016/j.stueduc.2016.08.007.

WCPS (Waterloo Course Perception Survey) Team. 2020. *Course Evaluation Project Pilot Test – Data Analysis Report*. Figure 29, page 48. Accessed 5 May 2022. https://uwaterloo.ca/teaching-assessment-processes/sites/default/files/uploads/documents/CEPT%20pilot%20test%20report_v12.pdf.

Wells, F. L. 1907. *A Statistical Study of Literary Merit: With Remarks on Some New Phases of the Method*. New York: Science Press. https://archive.org/details/statisticalstudy00well/page/n5/mode/2up.

Wikipedia. 2021. "Begging the Question." https://en.wikipedia.org/wiki/Begging_the_question.

Williams, W. M., and S. J. Ceci. 1997. "How'm I Doing? Problems with Student Ratings of Instructors and Courses." *Change: The Magazine of Higher Learning* 29 (5): 12–23. doi:10.1080/00091389709602331.

Zumbach, J., and J. Funke. 2014. "Influences of Mood on Academic Course Evaluations." *Practical Assessment, Research and Evaluation* 19 (4): 1–12. http://pareonline.net/getvn.asp?v=19&n=4.

Use of Non-Orthogonal Factor Analysis for Gauging Illusory Halo: A Technical Report[1]

John L. Michela, Ph.D.[2]

Gauging the extent of illusory versus true halo in students' evaluation of teaching (SET) surveys is difficult for various reasons. This technical report focuses on aspects of the survey items themselves as a difficulty, and it offers an approach to addressing this difficulty. The full context for this report is provided in an article by Michela (2022) written in reply to Cannon and Cipriani (2022), The present approach differs from approaches to gauging SET halo in Cannon and Cipriani (2022), and it yields different conclusions.

SET surveys differ from one another in (a) their range of areas of instruction and student experience that are covered, and (b) their extent of item similarity within each area. Examples of "areas" covered on SET surveys include perceived quality of explanations and other aspects of oral presentations by the instructor; perceived quality of teaching materials such as texts or problem sets; and perceived fairness and considerateness of the instructor toward students.

At the level of analysis of the survey items, intercorrelations among items that are obtained with a given survey can be noticeably low when a wide range of areas is asked about, each with a single survey item. These low correlations may, however, not signal low halo, because low reliability of measurement for each of the areas assessed would suppress the magnitudes of correlations. A different survey could yield high intercorrelations when few areas are covered with multiple, similar, or overlapping survey items within each area. In this instance, halo could be low in a conceptual sense, with the correlations being high because of redundant content of items—which is not the same as halo. A halo effect is present when survey respondents' SET ratings are consistent with one another to an *unwarranted* extent. High correlations are warranted when item content is substantially similar. High correlations are also warranted when the different areas queried are truly consistent, being either jointly favourable or unfavourable across the course offerings being rated for favourability. High correlations are *not* warranted

[1] Deposited May 27, 2022, to the UWSpace document repository at the University of Waterloo.
This document also serves as Supplement one to Michela (2022).
[2] Department of Psychology, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada; jmichela@uwaterloo.ca

when, in fact, instructor attributes or behaviours in the different areas do not have consistent standings with one another, but they are rated as though they are consistent with one another.

One approach to overcoming this effect of survey design is to attempt to shift the level of analysis from the level of the survey item to the level of broader, latent concepts that raters use when answering the particular survey items available. As a prime example, students generally are attuned to whether instructors come across as considerate and caring. Students' conception of this aspect of course experience constitutes a psychological construct. Some surveys may ask multiple related questions and produce a factor in factor analysis that isolates this construct. Such a factor is illustrated in the second of the three factors reported in Table C in this document. Other surveys may have only a single question or no questions on this or any given topic.

By subjecting a SET survey to factor analysis with *non-orthogonal* rotation (i.e., allowing *correlated* factors), it is possible to identify latent constructs for the survey items and thus to begin to quantify the intercorrelations of the constructs themselves. In contrast to survey *items*, which potentially have considerable overlap (as when multiple items ask about similar topics), constructs corresponding to *factors* are understood to be at least somewhat distinct (or else distinct factors would not have emerged). Thus, in this approach, illusory halo is implied when *factor* correlations are quite high, yet there is little reason to believe that the constructs that correspond with the factors would actually be consistent with one another or co-occur to the extent of the correlations obtained.

Thus, use of factor analysis with non-orthogonal rotation promotes comparison of the more meaningful, construct-level correlations for the varying constructs addressed on various SET surveys, administered in different times and places. Factor analyses of three SET surveys, spanning multiple decades and countries, appear in Tables A through C of this report. (All of these tables are based on publicly available data as per the reference citation given with each table.) Because factor correlations are correlations between latent variables, we gain the benefit of eliminating measurement error as a contaminant of the levels of correlations obtained between factors. With this error removed, the following factor analyses reveal notably and consistently high correlations among SET-related constructs. These constructs are labeled at the bottom right of each table (under "Factor Content"). Readers may surmise that some of these correlations imply the operation of considerable illusory halo, given the nature of the constructs involved and the magnitudes of the correlations.

For example, in Table A, survey items mentioning learning objectives are dominant in the first factor, and this factor has a high correlation with the broader, second factor. However, there is little reason to believe that in most courses the students are sufficiently attentive to learning objectives, per se, to allow for the high correlation of 0.73 between the two factors mainly on the basis of *actual* co-occurrence of the matters of Factor 1 and Factor 2. Illusory halo is a prime candidate for explaining this high correlation.

In Table A it is also noteworthy that the survey item for "Grades returned in reasonable time" does not have a high loading on either of the two factors. This item is especially "concrete" or explicit in what it asks. According to literature cited in Michela (2022), such concreteness can be expected to *reduce* illusory halo—which is precisely what appears to have occurred with this survey item in relation to the others.

In Table B a very high correlation of 0.80 is seen between a factor for Course content preference (Factor 2) and one for Instructor performance and consideration for students (Factor 3). If these inferred labels for the factors capture the psychological constructs that governed students' responses, then there is no apparent basis for such a high correlation between these constructs other than illusory halo. Why would instruction truly be superior in courses that match students' preferences?

As a challenge to this halo-based interpretation, perhaps Factor 2 is more centrally concerned with students' perceptions of extent of learning in the rated course. If so, initially it may seem warranted for students to have been highly consistent (again, at $r = 0.80$) in their ratings for this factor for learning and for the factor centered on instructor performance (Factor 3). That is, this consistency (correlation) in ratings would be warranted if better instructors produce better learning, and if students are accurate in their perceptions of superior instruction and of learning.

However, the literature points strongly, instead, to a halo-based interpretation, because students are *not* reliably accurate in their perceptions of learning. In Deslauriers, McCarty, Miller, Callaghan, & Kestin (2019), students rated their perceived learning as relatively low under conditions of having received instruction that was, in reality, superior, both in terms of use of instructional methods that are favoured by educational experts and in terms of the greater *actual* learning that these methods generated in this study (as assessed on examinations given to both the treatment and control groups in the study). Carpenter, Witherby and Tauber (2020)

describe instructional components that lead students to believe that they have received superior instruction and learning, even though empirical studies do not support their consistent effectiveness. These components include instructor behavioural fluency (upright posture, vocal inflections, etc.) and "decorative" use of visual aids. Carpenter et al. (2020) state: "The appearance of clarity, organization, and visual representations can sometimes mislead students into thinking they have learned more than they actually have" (p. 139). Bjork, Dunlosky, and Kornell (2013) provide additional analysis and evidence of errors and illusions in perception of one's own learning.

As an aside, these lines of analysis and evidence refute some other studies' use of students' ratings of their perceived learning as a criterion for claiming validity of SET survey ratings. For example, the University of Toronto (Centre for Teaching Support & Innovation, 2018) reported an association of a composite SET score (incorporating several conventional SET survey items) with a survey item for perceived learning, under the heading "construct validity." If students are not very good at assessing their own learning, what else can the extremely high correlation (r = 0.94) between perceived learning and the SET composite score reflect, other than a form of halo[3]? Correlations between the SET composite score and several other variables, taken to be validating, similarly are so high that some form of halo may be dominant. These variables include "students' perceptions that the course was intellectually engaging (r = 0.86), students' levels of interest after taking the course (r = 0.91), students' willingness to recommend the course to others (r = 0.88), and whether the instructor generated enthusiasm for the topic (r = 0.79)" (p. 23). Moreover, the intercorrelations among the five survey items of this SET composite score also were high to an extent that raises concern about halo dominance. The median among these correlations was 0.80. Corresponding median correlations in the literature often are considerably lower, such as the median of 0.66 in Feistauer and Richter (2017), which the present author regards as typical (though still high enough for considerable operation of halo).

Returning to the present concern with factor analysis as a tool for gauging halo, Table C is included to further acknowledge that high correlations among factors can have other bases besides illusory halo. In particular, the correlations of the third, global evaluation factor with the first two factors have a proper logic to them. That is, global evaluation logically depends on evaluations of more specific components.

[3] See Fisicaro and Lance (1990) concerning some forms of halo.

However, it is difficult to judge whether the high correlation (r = 0.67) between the first two factors here is primarily from illusory halo. Substantial *actual* consistency (and thus true halo, not illusory halo) for instructors in their instructional quality (Factor 1) and in their attentiveness and fairness (Factor 2) is conceivable. One such possibility is that instructors who are highly motivated to perform well, *as* instructors, tend to do both of two things: Teach in ways that students perceive as particularly effective, and engage in other behaviours that induce perceptions of attentiveness and fairness. On the other hand, even though each factor has a recognizable theme, the range of instructor attributes or behaviours, at the item level, is rather broad within each factor. This breadth suggests that some amount of illusory halo operated in generation of the responses behind Table C.

As discussed further in Michela (2022), it is impressive to have seen rather similar factor intercorrelations across the various factor solutions here. These factor solutions are based on data collected in different decades, in different countries, and on different survey instruments. Admittedly, three factor solutions (or four, including those in Cannon and Cipriani's paper) are too few to establish that factor analysis with non-orthogonal rotation is generally valuable for gauging extent of halo in SET survey responses. Additional archival and original data should be viewed with this factor analytic lens. Nevertheless, the findings in this technical report and other findings in Michela (2022) suggest presence of considerable illusory halo in SET survey responses. Michela (2022) argues, contrary to Cannon and Cipriani (2022) and some others, that halo in SET is indicative of poor validity of SET, which is to say, poor fitness for use as a measure of teaching effectiveness.

References

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417-444. https://doi.org/10.1146/annurev-psych-113011-143823

Cannon, E., & Cipriani, G. P. (2022). Quantifying halo effects in students' evaluation of teaching. *Assessment and Evaluation in Higher Education*, *47* (1), 1-14. https://doi.org/10.1080/02602938.2021.1888868

Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On students' (mis)judgments of learning and teaching effectiveness: Where we stand and how to move forward. *Journal of Applied Research in Memory and Cognition, 9*(2), 137–151. https://doi.org/10.1016/j.jarmac.2019.12.009

Centre for Teaching Support & Innovation. (2018). *University of Toronto's Cascaded Course Evaluation Framework: Validation Study of the Institutional Composite Mean (ICM).* Toronto, ON: Centre for Teaching Support & Innovation, University of Toronto. https://teaching.utoronto.ca/wp-content/uploads/2018/09/Validation-Study_CTSI-September-2018.pdf

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, *116*(39), 19251-19257. https://doi.org/10.1073/pnas.1821936116

Feistauer, D., & Richter, T. (2017). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *42*(8), 1263-1279.

Fisicaro, S.A., and C. E. Lance. 1990. Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement* 14 (4): 419-429. https://doi.org/10.1177/014662169001400407

Michela, J. L. (2022)[4]. Toward understanding and quantifying halo in students' evaluation of teaching. *Assessment and Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2022.2086965

[4] The 2022 date refers to the year of on-line publication, on June 20; the eventual print version could turn out to have a later year of publication. The https://doi.org identification should remain stable over time.

Table A. Rotated Factor Loadings (Pattern Matrix) and Factor Correlations from Promax Oblique Rotation of Factors Derived from WCPS (2021) Correlation Table

Item Number	*F1*	*F2*	*Item Content*
Q1	***0.50***	0.28	Instructor identified the LOs
Q2	***0.97***	-0.10	LOs assessed through graded work
Q3	***0.52***	0.31	Activities prepared me for graded work
Q4	0.32	0.20	Grades returned in reasonable time
Q5	0.10	***0.80***	Instructor conveyed course concepts
Q6	0.12	***0.74***	Supportive environment helped me learn
Q7	-0.05	***0.91***	Instructor stimulated interest
Q8	0.02	***0.89***	Overall I learned a great deal
Q9	0.10	***0.83***	Overall learning experience excellent

Factor Correlations			*Factor Content*
F1	1.00		Learning objectives (LOs) and grading
F2	0.73	1.00	Instructor and course global evaluation
	F1	*F2*	

Note. The first three initial eigenvalues were 5.835, 0.807, and 0.670, Two factors were retained because the rotated solution with three factors showed no variables with pattern matrix loadings above the conventionally required value of 0.40. After the promax rotation of two factors, sums of squared factor loadings in the structure matrix were 5.281 and 4.290. Factors are shown in reverse order relative to these sums of squares to provide alignment of the factors with the order of survey items.

Source: Waterloo Course Perception Survey (WCPS) Team. (2020). *Course Evaluation Project Pilot Test — Data Analysis Report*. Figure 29, page 48. Accessed May 5, 2022, from https://uwaterloo.ca/teaching-assessment-processes/sites/default/files/uploads/documents/CEPT%20pilot%20test%20report_v12.pdf.

Table B. Rotated Factor Loadings (Pattern Matrix) and Factor Correlations from Promax Oblique Rotation of Factors Derived from Gillmore (1975)

Item Number	*F1*	*F2*	*F3*	*Item Content*
Q1	0.01	***0.77***	0.19	Course content
Q2	-0.02	-0.01	***1.02***	Instructor contribution
Q3	0.05	0.03	***0.90***	Instructor effectiveness
Q4	0.15	0.21	***0.56***	Use of class time
Q5	0.32	0.17	***0.45***	Instr. interest student learning
Q6	0.04	***0.78***	0.18	Amount learned
Q7	0.01	***0.99***	-0.11	Relevance and usefulness
Q8	***0.85***	0.09	-0.01	Evaluation techniques
Q9	***0.86***	-0.04	0.01	Reasonableness of workload
Q10	***0.81***	0.02	0.08	Clarity of requirements

Factor Correlations				*Factor Content*
F1	1.00			Course requirements
F2	0.74	1.00		Course content preference
F3	0.78	0.80	1.00	Instr. perf. and consideration
	F1	*F2*	*F3*	

Note. The first three initial eigenvalues were 7.596, 0.747, and 0.480. After the promax rotation of three factors, sums of squared factor loadings in the structure matrix were 6.192, 6.327, and 6.635. The three-factor solution has been selected here for its interpretability and comprehensiveness (inasmuch as three factors cumulatively explain 88% of matrix variance).

Source: Correlation Table 8 (p. 18), with exclusion of the first variable for the course overall, in: Gillmore, G. M. (1975). *Statistical analysis of the data from the first year of use of the student rating forms of the University of Washington instructional assessment system*. University of Washington: Educational Assessment Center. Accessed September 6, 2021, from https://files.eric.ed.gov/fulltext/ED118580.pdf.

Table C. Factor Loadings (Pattern Matrix) and Factor Correlations from Promax Non-Orthogonal Rotation of SET Data of UCLA Demonstration of Factor Analysis

Item Number	*F1*	*F2*	*F3*	*Item Content*
13	***.90***			Instructor well prepared
14	***.83***			Instructor scholarly grasp
15	***.73***			Instructor confidence
16	***.63***			Instructor focus lectures
17	***.52***			Instructor uses clear relevant examples
18		***.81***		Instructor sensitive to students
19		***.88***		Instructor allows me to ask questions
20		***.59***		Instructor is accessible to students outside class
21		***.44***		Instructor aware of students understanding
22		***.55***		I am satisfied with student performance evaluation
23			***.78***	Compared to other instructors, this instructor is
24			***.82***	Compared to other courses, this course is

Factor Correlations				*Factor Content*
F1	1.00			Perceived instructional quality
F2	0.67	1.00		Perceived instructor attentiveness and fairness
F3	0.75	0.73	1.00	Overall evaluation of instructor and course
	F1	*F2*	*F3*	

Note. The first three initial eigenvalues from principal axis factoring were 6.249, 1.229, and 0.719. Rotation was by promax with Kaiser normalization. Other details for this promax rotated solution were not provided in the source document.

Source: UCLA Statistical Consulting Service. (2021). *Factor Analysis: SPSS Annotated Output*. Accessed September 6, 2021, from https://stats.idre.ucla.edu/spss/output/factor-analysis/.

Equations for Deriving Effect Sizes for Individual Predictors and Sets of Predictors under Specified Conditions in Multiple Regression Analysis: A Technical Report[1]

John L. Michela, Ph.D.[2]

Many published papers do not fully report effect sizes of predictor variables' influences on outcome variable scores. For example, many papers' multiple regression (MR) analyses report regression coefficients for individual predictors without providing each predictor's unique proportion of variance explained. This unique proportion may be obtained by squaring semi-partial correlation (*sr*) values that are available as an option in many statistical analysis programs.

This incompleteness of information occurs in many empirical reports concerning students' evaluations of teaching (SET). For example, in an MR analysis by Cannon and Cipriani (2022), some of the predictor variables were included so that they could serve as validating criteria for the SET rating response, and other variables were included to document distortion in SET responses due to halo bias. In an article in reply to Cannon and Cipriani, Michela (2022) argued that it is important to isolate the effect sizes of as many of the validating and non-validating predictors as possible, so that the extent of halo in SET response—the focus of the research by Cannon and Cipriani—could best be gauged.

Accordingly, this document provides equations for deriving effect sizes for individual predictors and sets of predictors under specified conditions in MR analyses. The effect size metric being obtained in each instance is either Pearson *r* or proportion of variance explained, expressed as R^2 or sr^2.

Some computations require availability of the overall *F*-ratio (i.e., for the complete MR equation). If only the overall R^2 of an MR analysis is provided, its corresponding *F*-ratio is obtained as:

$$F = ((N - k - 1) * R^2) / (k * (1 - R^2)). \quad (1)$$

Here, N is the total number of survey respondents, and k is the total number of predictor variables in the overall MR equation.

---

[1] Deposited May 27, 2022, to the UWSpace document repository at the University of Waterloo.
This document also serves as Supplement two to Michela (2022).

[2] Department of Psychology, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada; jmichela@uwaterloo.ca

By algebraic transformation of equation (1), R^2 may be obtained from F as follows:

$$R^2 = (F * k) / ((N - 1) + (k * (F - 1))). \quad (2)$$

The adjusted R^2 provides a correction for the overfitting that occurs in sample data, thus providing a population estimate of R^2. It is calculated as:

$$R^2\_\text{adj} = 1 - (1 - R^2) * ((N - 1) / (N - k - 1)). \quad (3)$$

If R^2_adj is available but R^2 is not, the latter is recoverable from this transformation of (3):

$$R^2 = 1 - (1 - R^2\_\text{adj}) * ((N - k - 1) / (N - 1)). \quad (4)$$

When individual predictors' unstandardized regression weights (b_i) are given along with the overall R^2 of the MR equation, their effect sizes are obtainable as sr^2 values from:

$$sr^2_i = (F_i * (1 - R^2)) / (N - k - 1). \quad (5)$$

For use in equation (5), F_i may be estimated from the square of the predictor's t-ratio, because

$$t^2 = F. \quad (6)$$

In practice, the table of coefficients from an MR analysis is likely to show some combination of regression coefficients (bs), standard errors (SEs) or t-ratios that is sufficient for obtaining the necessary quantities to insert into equation (5).

A Pearson correlation, r, generally is recoverable from an MR analysis when there are two linear predictors (as in the case of Model 1 in Cannon and Cipriani's paper). Given an overall R^2, and having calculated sr^2_2 (that is, sr^2 for the second of two predictors),

$$r_1 = \text{SQRT}(R^2 - sr^2_2), \quad (7)$$

and vice versa for r_2 and sr^2_1.

The unique R^2 for a set of predictors (R^2_change) is recoverable when the F-ratio unique to that set (F_change) has been provided along with R^2 for the full equation (R^2_full):

$$R^2\_\text{change} = (F\_\text{change} * k\_\text{change} * (1 - R^2\_\text{full})) / (N - k\_\text{full} - 1). \quad (8)$$

Equation (8) is an algebraic transformation of a conventional equation for the F-ratio (F_change) concerning addition or subtraction of a set of k predictors (k_change) in an MR model:

$$F = ((N - k\_\text{full} - 1) * R^2\_\text{change}) / (k\_\text{change} * (1 - R^2\_\text{full})). \quad (9)$$

References and Sources

Cannon, E., & Cipriani, G. P. (2022). Quantifying halo effects in students' evaluation of teaching. *Assessment and Evaluation in Higher Education, 47* (1), 1-14. DOI:10.1080/02602938.2021.1888868

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kerlinger, F. N. & Pedhazur, E. J., (1973). *Multiple regression in behavioral research*. New York, NY: Holt Rinehart & Winston.

Michela, J. L. (2022)[3]. Toward understanding and quantifying halo in students' evaluation of teaching. *Assessment and Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2022.2086965

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

[3] The 2022 date refers to the year of on-line publication, on June 20; the eventual print version could turn out to have a later year of publication. The https://doi.org identification should remain stable over time.