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 (in press at the time of on-line posting of this report), 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 1 to Michela (in press).
[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 (in press), 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 (in press), 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 (in press) suggest presence of considerable illusory halo in SET survey responses. Michela (in press) 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. (in press). Toward understanding and quantifying halo in students' evaluation of teaching. *Assessment and Evaluation in Higher Education*.

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/.