

Are course evaluations subject to a halo effect?

Jenny A. Darby Loughborough University

Student evaluations of courses using scaled responses are commonly used in school by OfStEd inspectors and in colleges and universities as part of their external grading process. Research studies of course evaluations have a long history. An early example by Bassin (1974) included a series of Likert scales concerned with five aspects of teaching. These were lecture quality, exam quality, text suitability, participation and consideration. It was found that instructors of quantitative courses received lower ratings than those of non-quantitative courses. Pohlmann (1975) evaluated five aspects of courses, namely an overall view of how good the course was, how interested the tutor was in the student, how difficult the student found the course, whether assignments were clearly marked and how good the tutors' actual presentation was. Pohlmann found undergraduate students' evaluations were better on elective than on required courses. This use of scales has continued, with Rae (1997, pp. 113–25) and Shevlin *et al.* (2000) recommending using structured scales. These researchers, in common with many others, assume the scales used are independent. This assumption is central to many course evaluations. Different aspects of courses are usually compared by looking at mean scores of scales which are thought to be independent. Sadly, these may have little real value if the individual scores which make up those means are a result of responses on one scale being influenced by those on another.

In an early review of the literature by Cohen (1981) there was evidence of some attempt to look at correlations between evaluation scales. This tended to be limited to very specific and predicted areas such as that between a favourable student rating of instructors' skills and the student having received better grades. The issue of the independence of scales in general has been neglected in more recent texts on course evaluation methodology (e.g. Holcomb, 1998; Rae, 2002; Salas *et al.*, 2003), also in the more broadly based research methodology texts (e.g. Fowler, 2002, Hayes, 2000, and Shaughnessey *et al.*, 2000).

The present study questions the independence of measures on evaluation scales in the light of the halo effect, which is well known in the field of person perception but is not a concept which has been commonly applied to course evaluation. Blum and Naylor (1968, p. 200) see the halo effect sim-

ply as the 'tendency to let our assessment of an individual on one trait influence our evaluation of that person on other specific traits'. This definition allows any influence to be positive or negative.

According to Thorndike (1920), one of the problems with trying to show whether a halo effect has occurred is that the various items may actually be related and so any relationship is based on real similarities rather than a social influence. It is not possible to totally eliminate this problem but as Thorndike (1920), and more recently Mi-Young and Jyotika (2003), acknowledge it is satisfactory if reasonable steps are taken to ensure there are differences between the items. A method of testing whether a halo effect occurs, which was originally used by Thorndike and Hagen (1977), involved correlating scores for various factors. This is one of the methods to be adopted in this present study. In addition this study looks not only at the influence of one set of Likert scales on another but also at any influence between different types of format. Typically many structured evaluation forms incorporate an open-ended section. Kobrynowicz and Biernat (1997) justify this by arguing that open-ended response forms allow a greater degree of expression than structured Likert-style response forms. These two evaluation formats are here to be examined in terms of the impact of the halo effect to see whether a pattern of responses occurs within one type of format and whether there is an influence across the two formats.

The three hypotheses to be tested are:

- 1 Evaluations on a structured questionnaire concerning different elements of a course would correlate positively, displaying a halo effect.
- 2 Responses on an open-ended section would correlate positively, displaying a halo effect.
- 3 There would be a positive correlation between evaluations on the Likert scale and the open-ended evaluation.

Method

Participants

Student course evaluations were obtained from 161 university lecturers attending thirteen different half-day probationary training courses on aspects of teaching. These included topics such as lecturing skills, working with groups and encouraging critical thinking. The students were lecturers from a wide range of disciplines from five different universities in the East Midlands. Only four lecturers attended more than one course. None of them attended more than two courses. This overlap in the sample was considered so small as not to have had an impact on the statistical analysis.

Evaluation questionnaire used

This was in common use but the individual statements were categorised so that comparisons could be made with any open-ended responses. The form

Table 1 Example of a Likert-style structured evaluation form

	<i>Very poor</i>	<i>Poor</i>	<i>Average</i>	<i>Good</i>	<i>Very good</i>
Consistency with publicity					
<i>Hygiene</i>					
Relevance to your needs					
<i>Content</i>					
Quality of presentations					
<i>Human-related</i>					
Quality of group management					
<i>Human-related</i>					
Quality of audio-visual					
<i>Hygiene</i>					
Quality of hand-out materials					
<i>Hygiene</i>					
Enjoyability of the course					
<i>Content</i>					
Usefulness of the course					
<i>Content</i>					
Integration of different components					
<i>Human-related</i>					
Appropriateness of level					
<i>Human-related</i>					
Followed good Equal Opportunities practice					
<i>Hygiene</i>					
Efficiency of course administration					
<i>Hygiene</i>					
Overall					
The best thing					
Another good thing					
The worst thing					
Another bad thing					

is shown in Table 1 together with the evaluation category groupings, which are in italics. The categories were not included in the copies given to the course participants. The categories were derived thematically using a hypothetical-deductive approach (Hayes, 2000, p. 179) and then an inductive approach. The three categories thus had their conceptual origins in a review of previous research studies and are as follows:

- 1 The category, for convenience in this article referred to as ‘human-related factors’, was based on work by, for example, Herzberg (1966), who pointed out how when people are feeling positive about their work they react favourably to their colleagues and others they can relate to. Another research study, by Parrot *et al.* (1988), stressed the tendency to react positively to persons, and Morgan *et al.* (1997) stressed the importance of groups and how we turn to them for support.

- 2 The category referred to as 'hygiene factors' was again based on work by Herzberg (1966), who highlighted the use of 'hygiene factors' when individuals want to express displeasure. These tended to be such things as working conditions and administrative items. Parrott *et al.* (1988) showed how individuals use inanimate areas to express negative views. In the present study these 'hygiene factors' include items such as joining instructions, teaching environment and visual aids.
- 3 The category referred to as 'content factors' included feelings about the content of the course which are to do with the reaction of the participants, for example whether they enjoyed it and felt it was useful. This is considered by Furedi (2003), among many other researchers, to be important as a factor to be included in course evaluations.

The twelve structured statements were classified into the three categories by five raters acting independently. Of the sixty statements included in this task fifty-seven were placed unanimously in the same categories by the raters. Conceptually the three categories were individually very different. The reliability of the three scales was shown to be acceptable when tested by means of a Cronbach alpha test, which showed a reading of 0.756 for the 'human-related' category, 0.582 for the 'hygiene' category and 0.843 for the 'content' category.

The statements made by the students on the open-ended evaluation forms were categorised by the researcher. Twenty per cent were selected at random by an assistant, who was instructed in the categorisation scheme. Completely independently a total of thirty-two forms were scored by this assistant. Forty-seven individual statements on these forms were placed in categories and forty-three were placed by the assistant in the same categories as by the researcher. This was a 91 per cent matching rate.

Numerical scoring of the open-ended evaluations

These were scored according to the order of the comments made. For each of the 'favourable' comments the first made was awarded a score of 4, the second 3, the third 2 and the fourth and subsequent comments 1. When no comment was made that category was awarded 0. The 'unfavourable' comments were scored separately but used the same numerical scale. This method of scoring took into account the order effect noted by Sherman and Klein (1994), Wyer *et al.* (1994) and Swann and Gill (1997), that the first thing said is the most important to the speaker.

Results

Hypothesis 1. Evaluations on a structured questionnaire concerning different elements of a course would correlate positively, displaying a halo effect

The hypothesis is supported, for, as can be seen in Table 2, the component matrix of a factor analysis shows all the components of the three categories

Table 2 Factor analysis using the extraction method component analysis with two components extracted

	<i>Component</i>	
	<i>1</i>	<i>2</i>
1 Consistency with publicity <i>Hygiene</i>	0.570	-0.054
2 Relevance to your needs <i>Feelings about content</i>	0.752	-0.406
3 Quality of presentations <i>Human-related</i>	0.689	-0.209
4 Quality of group management <i>Human-related</i>	0.590	0.095
5 Quality of audio-visual <i>Hygiene</i>	0.443	-0.176
6 Quality of hand-out materials <i>Hygiene</i>	0.554	-0.191
7 Enjoyability of the course <i>Feelings about course</i>	0.805	-0.217
8 Usefulness of the course <i>Feelings about course</i>	0.742	-0.345
9 Integration of different components <i>Human-related</i>	0.670	0.338
10 Appropriateness of level <i>Human-related</i>	0.796	0.133
11 Followed good Equal Opportunities practice <i>Hygiene</i>	0.489	0.633
12 Efficiency of course administration <i>Hygiene</i>	0.522	0.605
Open-ended <i>Best 'Human-related'</i>	0.289	0.166
Open-ended <i>Best 'Hygiene'</i>	-0.284	-0.146
Open-ended <i>Best 'Content'</i>	0.136	-0.059

on the Likert scales listed 1–12 in the table are heavily loaded on the first factor.

In order to compare the three categories themselves on the Likert scale a Pearson product moment correlation coefficient was carried out on a combined score for each of the sub-scales within each category (Table 3) and indicates a high positive correlation between each pair of the three categories. According to Sheehan and DuPrey (1999) correlations at these levels form a meaningful relationship between the factors. It confirms the halo effect has an impact on the way in which evaluation forms are completed.

The importance of this result is that the correlation takes into account the full range of opinions of the students, from those who are reacting extremely

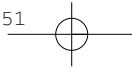


Table 3 Correlation for positive scores of individuals for the three factors on the Likert structured evaluations

<i>Elements of course</i>	<i>Human-related</i>	<i>Hygiene</i>	<i>Content</i>
Human-related	1.00 (0.000)	0.612 (0.000)	0.666 (0.000)
Hygiene	0.612 (0.000)	1.00 (0.000)	0.600 (0.000)
Content	0.666 (0.000)	0.600 (0.000)	0.1.00 (0.000)

Notes $n = 161$. Significance in brackets.

favourably to those who are reacting far less favourably. The three categories each measure very different factors, as is evidenced by the ease with which the conceptual grouping of items was originally carried out. If the unreliability of the three scales is taken into account using the Cronbach alpha readings and the correction for attenuation (which takes account of scale unreliability when testing for a correlation) is calculated the correlations increase considerably for all pairs of comparisons. Between the 'human-related' and 'hygiene' scales the correlation is 0.92, between the 'human-related' and 'content' scales the correlation is 0.835 and between the 'hygiene' and 'content' scales the correlation is 0.94. The individuals are reacting to all three categories in a very similar manner, whether favourably or unfavourably.

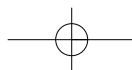
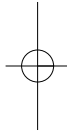
Hypothesis 2. Responses on an open-ended section would correlate positively, displaying a halo effect

The hypothesis is not supported, for, as can be seen in Figure 1, when a multiple correlation is carried out between the positive reactions to the three main elements on the open-ended section of the questionnaire all are below a level which, according to Sheehan and Duprey (1999), would indicate a meaningful relationship.

Further, if the negative comments are considered, there are also no significant correlations. Those who dislike one aspect of a course do not necessarily dislike another. There is no evidence of a halo effect between the unfavourable responses on the open-ended sections of the questionnaire.

Hypothesis 3. There would be a positive correlation between evaluations on the Likert scale and the open-ended evaluation

As can be seen in Table 2 the factor analysis does not provide any support for the hypothesis. The three items referring to the open-ended responses at the very bottom of the table, namely Best 'human-related', 'hygiene' and 'content', do not load heavily on the first factor as do the Likert scales. Furthermore the correlations, shown in Table 4, between the Likert scales and the open-ended scales are all so low as to show the course participants' responses on the structured questionnaire, and the open-ended sections bear little rela-



	<i>Best human-related</i>	<i>Best hygiene</i>	<i>Good content</i>	<i>Worst human-related</i>	<i>Worst hygiene</i>	<i>Bad content</i>
<i>Best human-related</i>	1.00	-0.038 (0.628)	-0.117 (0.139)	0.060 (0.447)	0.149 (0.059)	0.061 (0.444)
<i>Best hygiene</i>	-0.038	1.00	-0.085 (0.282)	0.215 (0.006)	0.211 (0.007)	0.248 (0.002)
<i>Good content</i>	-0.117	-0.086	1.00	0.108 (0.173)	0.088 (0.268)	0.069 (0.382)
<i>Worst human-related</i>	0.060	0.215	0.108	1.00	0.038	0.032 (0.690)
<i>Worst hygiene</i>	0.149	0.211	0.088	0.038	1.00	-0.064 (0.422)
<i>Bad content</i>	0.061	0.248	0.069	0.032	0.064	1.00

Figure 1 Open-ended correlations for individuals. (Significance in brackets)

Table 4 Likert structured and open ended correlations for individuals

	<i>Open-ended</i>					
	<i>Best human-related</i>	<i>Best hygiene</i>	<i>Good content</i>	<i>Worst human-related</i>	<i>Worst hygiene</i>	<i>Bad content</i>
Human-related (Likert)	0.227 (0.004)	-0.201 (0.011)	0.027 (0.738)	-0.238 (0.002)	-0.214 (0.006)	-0.033 (0.682)
Hygiene (Likert)	0.201	-0.165 (0.036)	0.108 (0.172)	-0.129 (0.102)	-0.200 (0.011)	-0.117 (0.139)
Feelings Content (Likert)	0.202 (0.010)	-0.242 (0.002)	0.129	-0.347 (0.000)	-0.105 (0.185)	-0.106 (0.179)

tion to each other. This really shows two things. First, that the two forms are being responded to differently and the halo effect does not cross the boundaries of the design of the evaluation forms. Second, and most important, it does suggest that with the Likert scales a halo effect is occurring. The evidence of the open-ended responses indicates the students do not regard all aspects equally favourably or unfavourably, as would be suggested if the Likert scales are taken at face value. This does suggest the correlations between measures on the Likert scales are more a result of a halo effect than of a genuine liking or disliking by individuals of the various measures.

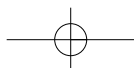
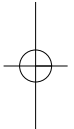


Discussion

The results of this study highlight three major characteristics of a type of pencil-and-paper evaluation form in common use. First, it appears, with a Likert-type scale, students who reportedly like one aspect of a course also appear to like another and those who reportedly dislike one aspect also appear to dislike others. Although the three categories used here are conceptually very different the factor analysis shows the twelve items which make up these three categories are heavily loaded on a single component. Furthermore, the three categories are shown to correlate highly and when the Cronbach alpha is used to provide a correction for attenuation the correlation is even more marked. It is argued in this article that a halo effect seems to have occurred. It appears that the halo effect can be moved out of the area of person perception and can be applied to Likert-style course evaluations. This interpretation is supported by the fact that this pattern of responses does not occur with open-ended evaluations. This would appear to indicate overall favourable or unfavourable response patterns on a Likert scale reflect a halo effect rather than student views. These results indicate the halo effect needs to be taken into account when considering the results of any course evaluation using a Likert-style structured scale. It should be stressed these findings occurred not with impressionable schoolchildren, or even older students, but with lecturers. It should also be stressed that the results are not simply a case of the courses being excellent ones and the students liking them. The correlation between categories of scales shows that students who respond favourably to one aspect also respond favourably to another. It also shows how the halo effect operates in reverse. Students who respond less favourably to one aspect also react less favourably to other aspects of the evaluation.

Second, it is noticeable the halo effect does not occur with open-ended evaluations. Students offer their views of a course in terms of unrelated statements. They may react favourably to one aspect of the course and unfavourably to another. There does not seem to be the same 'mental set' when it comes to filling in this type of evaluation. It would appear Kobrynowicz and Biernat's (1997) argument that open-ended response forms allow for a greater degree of expression than structured Likert style response forms can be developed a little further. Not only are students freer to express themselves but also their choice of response is not influenced by the constraints or influences of the halo effect which appears to restrict responses on Likert scales.

Third, there would not appear to be a link between the responses on the Likert scales and the open-ended responses. Students on the courses who, on the rating scales, say they like the presenters do not necessarily say they like the presenters when they give open-ended responses. It would seem students react differently to different styles of evaluation forms. The failure to identify a halo effect with the open-ended responses suggests responses on the Likert scales are subject to very different influences from those on the open-



ended evaluation forms.

This study has implications for those using Likert-type scales for evaluating courses. It would appear individual scales are not regarded independently by students, for, probably unknowingly, a halo effect occurs and their feelings about one aspect of a course would seem to influence their expressed views of other aspects. Further, the fact there seems to be no connection between Likert scales and open-ended comments would suggest students are responding very differently to the two formats. Interpretations based on evaluation forms in common use in schools by OfStEd, and also those used in colleges and universities, need to take into account the findings about the relationship noted here between scales on an evaluation form.

References

- Bassin, W. M. (1974), 'A note on the biases in students' evaluations of instructors', *Journal of Experimental Education* 43, 16–17.
- Blum, M. I., and Naylor, J. C. (1968), *Industrial Psychology: its Theoretical and Social Foundations*. New York: Harper & Row.
- Cohen, P. A. (1981), 'Student ratings of instruction and student achievement: a meta-analysis of multi-section validity studies', *Review of Educational Research* 51 (3), 281–309.
- Fowler, F. J. (2002), *Survey Research Methods*, London: Sage.
- Furedi, F. (2003), 'Students are not customers', *Outlook* 226, 13.
- Hayes, N. (2000), *Doing Psychological Research*, Buckingham and Philadelphia: Open University Press.
- Herzberg, F. (1966), *Work and the Nature of Man*, Cleveland OH: World Publishing.
- Holcomb, J. (1998), *Training Evaluation made Easy: Making your Training worth every Penny*, London: Kogan Page.
- Kobryniewicz, D., and Biernat, M. (1997), 'Decoding subjective evaluations: how stereotypes provide shifting standards', *Journal of Experimental Social Psychology* 33, 579–601.
- Mi-Young, O., and Jyotika, R. (2003), 'Halo-effect: conceptual definition and empirical exploration with regard to South Korean subsidiaries of US and Japanese multinational corporations', *Journal of Communication Management* 7 (4), 317–30.
- Morgan, D., Carder, P., and Neal, M. (1997), 'Are some relationships more useful than others? The value of similar others in the networks of recent widows', *Journal of Social and Personal Relationships* 14, 745–59.
- Parrot, W. G., Sabini, J., and Silver, M. (1988), 'The roles of self-esteem and social interaction in embarrassment', *Personality and Social Psychology Bulletin* 14, 191–202.
- Pohlmann, J. T. (1975), 'A multivariate analysis of selected class characteristics and student ratings of instructions', *Multivariate Behavioural Research* 10 (1), 81–91.
- Rae, L. (1997), *How to Measure Training Effectiveness* (third edition), Aldershot: Gower Publishing.
- Rae, L. (2002), *Assessing the Value of your Training: the Evaluation Process from Training Needs to the Report to the Board*, Aldershot: Gower Publishing.
- Salas, E., Milham, L. M., and Bowers, C. A. (2003), 'Training evaluation in the military: misconceptions, opportunities and challenges', *Military Psychology* 15 (1), 3–16.
- Shaughnessy, J. J., Zechmeister, E. B., and Zechmeister, J. S. (2000), *Research Methods in Psychology* (fifth edition), Boston MA: McGraw-Hill.
- Sheehan, E. P., and DuPrey, T. (1999), 'Student evaluations of university teaching', *Journal of Institutional Psychology* 26 (3), 188–93.
- Sherman, J. W., and Klein, S. B. (1994), 'Development and representation of per-

- sonality impressions', *Journal of Personality and Social Psychology* 67, 972–83.
- Shevlin, M., Banyard, P., Davies, M., and Griffiths, M. (2000), 'The validity of student evaluation of teaching in higher education: love me, love my lectures?' *Assessment and Evaluation in Higher Education* 25 (4), 397–405.
- Swann, W. B., Jr, and Gill, M. J. (1997), 'Confidence and accuracy in person perception: do we know what we think we know about our relationship partners?' *Journal of Personality and Social Psychology* 73, 747–57.
- Thorndike, E. L. (1920), 'A constant error in psychological ratings', *Journal of Applied Psychology* 4, 25–9.
- Thorndike, E. L., and Hagen, E. (1977), *Measurement and Evaluation in Psychology and Education* (second edition), New York: Wiley.
- Wyer, R. S., Jr, Budenheim, T. I., Lambert, A. J., and Swan, S. (1994), 'Person perception judgement: pragmatic influences on impressions formed in a social context', *Journal of Personality and Social Psychology* 66, 254–67.

Acknowledgement

This study is based on the author's doctoral thesis.

Address for correspondence

Department of Social Sciences, Loughborough University, Loughborough, Leicestershire LE11 3TU. *E-mail* j.a.darby@lboro.ac.uk

Copyright of Research in Education is the property of Manchester University Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.