Are course evaluations subject to a 'halo effect'?

Research in Education 2007 No. 77. 46-55

ARE COURSE EVALUATIONS SUBJECT TO A 'HALO EFFECT?

Jenny A. Darby PhD.
Department of Social Sciences, Loughborough University, Loughborough, Leicestershire. LE11 3TU
E mail J.A.Darby@lboro.ac.uk
Keywords: Course**,** Evaluation, Halo effect, Likert scale

Abstract:

Many course evaluations including those used in schools by OFSTED, and colleges and universities employ a number of scales as a means of evaluating various aspects of the educational experience of the student. It tends to be assumed students consider the scales independently. This paper argues students are influenced by a 'halo effect' when making judgements about the merits of aspects of a course. In this study student evaluations were obtained from 161 university lecturers attending probationary training courses. With a Likert style structured scale a favourable evaluation on one aspect of a course was shown to be linked with favourable evaluations on other aspects. Similarly with unfavourable evaluations different aspects were shown to correlate. This was not the case with an open ended evaluation format. There were no links between reactions on the two types of evaluations. Implications for the interpretation of course evaluations are discussed.

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 student's evaluations were better on elective than required courses. This use of scales has continued with Rae (1997 p 113-125), 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 (eg Holcomb 1998, Rae 2002, and Salas et al 2003), also in the more broadly based research methodology texts (eg. 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 evaluations. Blum and Naylor (1968 p. 200) see the 'halo effect' simply 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 for any influence to be positive or negative.

One of the problems with trying to show whether a 'halo effect' has occurred is, according to Thorndike (1920), 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 not only looks at the influence of one set of Likert scales on another but also any influence between different types of format. Typically many structured evaluation forms incorporate an open ended section. Kobrunowicz and Biernat (1997) justify this by arguing that open ended response forms allow for 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:
Hypothesis 1*: Evaluations on a structured questionnaire concerning different elements of a course would correlate positively displaying a 'halo effect'*.
Hypothesis 2. *Responses on an open ended section would correlate positively displaying a 'halo effect'*.
Hypothesis 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 these 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 usage but the individual statements were categorised so that comparisons could be made with any open ended responses. The form 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 thermatically using a hypothetic-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 paper 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 things such 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), amongst 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 60 statements included in this task 57 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 Chronbach alpha test which showed a reading of .756 for the 'human related' category; .582 for the 'hygiene' category and .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. Totally independently a total of 32 forms were scored by this assistant. 47 individual statements on these forms were placed in categories and 43 were placed by the assistant in the same categories as the researcher. This was a 91% matching rate.

Table 1
Example of a Likert style structured evaluation form

	Very Poor	Poor	Average	good	Very good
Consistency with publicity *Hygiene*					
Relevance to your needs *Content*					
Quality of presentations *Human related*					
Quality of group management *Human related*					
Quality of audio-visual *Hygiene*					
Quality of handout materials *Hygiene*					
Enjoyability of the course *Content*					
Usefulness of the course *Content*					
Integration of different components *Human related*					
Appropriateness of level *Human related*					
Followed good equal opportunities practice *Hygiene*					
Efficiency of course administration *Hygiene*					
Overall					
The best thing					
Another good thing					
The worst thing					
Another bad thing					

Numeric 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 four the second three, the third two, and the fourth and subsequent comments one. When no comment was made that category was awarded zero. 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 on the Likert scales listed one to twelve in the table are heavily loaded on the first factor.

Table 2
Showing factor analysis using the extraction method component analysis with two components extracted.

	Compnent 1	Component 2
1. Consistency with publicity *Hygiene*	**.570**	**-.054**
2. Relevance to your needs *Feelings about content*	**.752**	**-.406**
3. Quality of presentations *Human related*	**.689**	**-.209**
4. Quality of group management *Human related*	**.590**	**.095**
5. Quality of audio-visual *Hygiene*	**.443**	**-.176**
6. Quality of handout materials *Hygiene*	**.554**	**-.191**
7. Enjoyability of the course *Feelings about course*	**.805**	**-.217**
8. Usefulness of the course *Feelings about course*	**.742**	**-.345**
9. Integration of different components *Human related*	**.670**	**.338**
10. Appropriateness of level *Human related*	**796**	**.133**
11. Followed good equal opportunities practice *Hygiene*	**.489**	**.633**
12. Efficiency of course administration *Hygiene*	**.522**	**.605**
Open ended Best 'Human related'	**.289**	**.166**
Open ended Best 'Hygiene'	**-.284**	**-.146**
Open ended Best 'Content'	**.136**	**-.059**

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.

Table 3
Showing correlation for positive scores of individuals for the three factors on the Likert structured evaluations. N=161. Significance in brackets

Elements of course	Human Related	Hygiene	Content
Human related	1.00 (.000)	.612 (.000)	.666 (.000)
Hygiene	.612 (.000)	1.00 (.000)	.600 (.000)
Content	.666 (.000)	.600 (.000)	.1.00 (.000)

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 favourably to those who are reacting far less favourably. The three categories each measure very different factors as is evidenced by the ease at which the conceptual grouping of items was originally carried out. If the unreliability of the three scales is taken into account using the Chronbach alpa 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 .92, between the 'human related' and 'content' scales the correlation is .835 and between the 'hygiene' and 'content' scales

the correlation is .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 table 4 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.

Table 4
Open ended correlations for individuals. Sig in brackets

	Best Hum related	Best Hygiene	Good Content	Worst Hum related	Worst Hygiene	Bad Content
Best Hum related	1.00	-.038 (.628)	-.117 (.139)	.060 (.447)	.149 (.059)	.061 (.444)
Best hygiene	-.038	1.00	-.085 (.282)	.215 (.006)	.211 (.007)	.248 (.002)
Good Content	-.117	-.086	1.00	.108 (.173)	.088 (.268)	.069 (.382)
Worst Hum related	.060	.215	.108	1.00	.038	.032 (.690)
Worst Hygiene	.149	.211	.088	.038	1.00	-.064 (.422)
Bad Content	.061	.248	.069	.032	.064	1.00

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 5, between the Likert scales and the open ended scales are all so low as to show the course participant's responses on the structured questionnaire and the open ended section bear little relation 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 importantly 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 a genuine liking or disliking by individuals of the various measures.

Table 5
Likert structured and open ended correlations for individuals

	Open ended Best Hum related	Open ended Best Hygiene	Open ended Good Content	Open ended Worst Hum related	Open ended Worst Hygiene	Open ended Bad Content
Human related (Likert)	.227 (.004)	-.201 (.011)	.027 (.738)	-.238 (.002)	-.214 (.006)	-.033 .682)
Hygiene (Likert)	.201	-.165 (.036)	.108 (.172)	-.129 (.102)	-.200 (.011)	-.117 (.139)
Feelings Content (Likert)	.202 (.010)	-.242 (.002)	.129	-.347 (.000)	-.105 (.185)	-.106 (.179)

DISCUSSION

The results of this study highlight three major characteristics of a type of pencil and paper evaluation forms 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 Chronbach alpha is used to provide a correction for attenuation the correlation is even more marked. It is argued in this paper 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' is one which 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 school children, 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 student 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 Kobrunowicz and Biernat (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 to 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.

Acknowledgement:

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

REFERENCES

Bassin, W. M. (1974) A note on the biases in students' evaluations of instructors'. *Journal of Experimental Education* 43. 16-17

Blum, M.l. and Naylor, J.c. (1968*) Industrial Psychology: Its theoretical and social foundations.* Harper and Row: New York

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 *Autlook* 226 13.

Hayes, N. (2000) *Doing psychological research*. Buckingham. Philadelphia. Open University Press

Herzberg, F. (1966) *Work and the nature of man*. Cleveland. World Pyublishing.

Holcomb, J. (1998*) Training evaluation made easy: Making your training worth every penny.* Kogan Page.

Kobrynowicz, 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. 317-330

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. 745-759

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) (3rd edition). *How to Measure Training Effectiveness*. Hampshire, England. Gower Publishing Ltd.

Rae, L. (2002) *Assessing the value of your training: The evaluation process from training needs to the report to the board*. Aldershot, UK: Gower Publishing Co.

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*. 5th edit. Boston: McGraw-Hill

Sheehan, E.P. and DuPrey, T. (1999) Student evaluations of university teaching. *Journal of Institutional Psychology.* 26.3. 188-193

Sherman,J.W. and Klein, S.B. (1994) Development and representation of personality impressions. *Journal of Personality and Social Psychology*. 67, 972-983

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. 747-757

Thorndike, E.L. (1920) A constant error in psychological ratings. *Journal of Applied Psychology*. 4. 25-29

Thorndike, E.L. and Hagen, E. (1977*) Measurement and evaluation in psychology and education.* (2nd Edit). Wiley: New York.

Wyer,R.S. Jr. Budesheim, 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. 254-267.