



Higher Education Research and Development Society of Australasia, Inc

The Student Experience

Proceedings of the

32nd HERDSA Annual Conference

6-9 July 2009

Darwin, Australia

Huybers, T. (2009) Best-Worst Scaling: a 'magnifying glass' method for university teaching evaluation, in *The Student Experience, Proceedings of the 32nd HERDSA Annual Conference, Darwin, 6-9 July 2009: pp 219-229.*

Published 2009 by the
Higher Education Research and Development Society of Australasia, Inc
PO Box 27, Milperra, NSW 2214, Australia
www.herdsa.org.au

ISSN: 0155 6223
ISBN: 0 908557 78 7

This research paper was reviewed using a double blind peer review process that meets DEEWR requirements. Two reviewers were appointed on the basis of their independence, expertise and experience and received the full paper devoid of the authors' names and institutions in order to ensure objectivity and anonymity. Where substantial differences existed between the two reviewers, a third reviewer was appointed. Papers were evaluated on the basis of originality, quality of academic merit, relevance to the conference theme and the standard of writing/presentation. Following review, this full paper was presented at the international conference.

Copyright© 2009 HERDSA and the authors. Apart from any fair dealing for the purposes of research or private study, criticism or review, as permitted under the Copyright, Design and Patent Act, 2005, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers at the address above.

Best-Worst Scaling: a 'magnifying glass' method for university teaching evaluation

Twan Huybers

UNSW@ADFA, Canberra, Australia
t.huybers@adfa.edu.au

Student evaluation of teaching at universities is subject to widespread debate. The discussion in the academic literature relates to various aspects of teaching evaluation including the range of determining factors and, broadly, the validity of its measurement of teaching effectiveness. This paper focuses on the measurement of the relative importance of the different aspects of teaching covered in a teaching evaluation instrument. Best-Worst Scaling is proposed as a new method to evaluate university teaching. Based on a pilot study of a small undergraduate class, the findings of the study demonstrate how Best-Worst Scaling, in comparison with conventional rating scales, better highlights the true differences in measurements across a range of teaching evaluation criteria. In doing so, it provides enhanced feedback to educators about students' perceived teaching performance along various dimensions of teaching.

Keywords: student evaluation of teaching, rating scales, Best-Worst Scaling

Introduction

Student evaluation of university teaching is a hotly debated subject of discussion around university campuses, in the popular press (Davies *et al.*, 2007; McDonald & Mills, 2007), and in the scholarly literature (see, for instance, Greenwald, 1997; Haskell 1997; Marsh & Roche 1997; Clayson & Sheffett, 2006). Sproule (2000) identifies two fundamental purposes of teaching evaluation results: the summative function with a view to management decisions about academic career progress, recruitment purposes and the like, and the formative function as a feedback tool for university teachers on how to improve the quality and effectiveness of their teaching and their courses.

The method and metrics underlying student evaluations of teaching have received criticism. Teaching evaluation results may be affected by factors such as grades, gender, and class size (see, for instance, Koh & Tan, 1997; Isely & Singh, 2005; Badri *et al.*, 2006). One particular aspect of criticism, which is addressed in this paper, is that of response style bias (Dolnicar & Gruen, 2007).

This paper concentrates on the 'pure' formative function of student evaluations of university teaching. University educators who are genuinely concerned about the quality of the education experience seek comments and constructive criticism with a view to improving their teaching. In addition to student evaluations, various complementary tools can be used for that purpose including self evaluation (for instance using reflective journals), and peer review of teaching materials and class teaching (Webster, 1990; Dochy *et al.*, 1999). The subject of investigation in the current study is the measurement of the relative importance of the various aspects of teaching that are contained in an instrument for student evaluation of teaching.

The paper is organised in the following way. The next section addresses the conventional approach to measuring student evaluations of the effectiveness of university teaching, and introduces Best-Worst Scaling as an alternative to the conventional method. Next, the details of the study's survey are provided. Subsequently, the survey results are presented and discussed followed by the concluding comments in the final section.

Teaching evaluation instruments and best-worst scaling

Student evaluations of university teaching have traditionally employed survey instruments that contain a number of questions in the form of statements for which students are asked to indicate their level of (dis)agreement. The scale that is typically used is of a Likert type with five or seven points ranging from "Strongly Agree" to "Strongly Disagree".

The use of Likert type scales is widespread despite concerns about its vulnerability in relation to response style bias (Baumgartner & Steenkamp, 2001). Response style bias refers to a respondent's consistent answering pattern irrespective of the actual questions asked. Baumgartner and Steenkamp (2001) identify seven response styles including acquiescence response style (agreement tendency), extreme response style (the systematic use of the endpoints of a scale) and midpoint responding. Alternative response formats include even-numbered Likert scales to deal with respondents' tendency towards neutral opinions (Onwuegbuzie & Weems, 2004), and binary response scales (Cronbach, 1950) to elicit even less ambiguous responses.

The application of conventional Likert type rating scales to teaching evaluations is expected to be prone to the above response style bias problems. For instance, certain students may exhibit a midpoint response style pattern while others may tend to give extreme, affirmative responses across the board. Either way, because of the potential lack of discrimination among the various teaching criteria, educators would receive little useful guidance on where to focus specific teaching enhancement efforts.

The pitfalls of applying conventional types of rating scales to teaching evaluations provided the impetus to exploring an alternative measurement method. The key condition of such an alternative method had to be its ability to identify clearly and validly the differences in student perceptions of the various teaching criteria contained in the evaluation instrument. If teaching evaluations are intended to provide guidance to educators about the effectiveness of the various features of their teaching activities, they ought to feel confident about the relative results for those features as derived from the evaluation exercise.

In this paper, Best-Worst Scaling (BWS) is proposed as an alternative method that enhances discrimination between various items on a list of items of concern. In a BWS experiment, respondents are asked to indicate the "best" or "most applicable" item and the "worst" or "least applicable" item in a number of experimentally designed sets of items each of which is a subset of a larger overall set. The experiment forces respondents to think about the relative importance of each item. Since this method captures respondents' maximum perceptual differences on an underlying latent scale, BWS is sometimes referred to as 'maximum difference scaling' (Cohen, 2003). Chrzan and Golovashkina (2006) report on a study which compares BWS and five other importance measures and find that BWS does better than the other measures on a number of different dimensions including importance differentiation and predictive validity. On the other hand, BWS is found to be the most time consuming method.

The conventional ratings task, while involving the shortest length of questionnaire completion time, was the inferior method on all other dimensions.

The earliest published application of BWS is Finn and Louviere (1992) in the context of the measurement of the level of concern about various aspects of food safety as well the comparison of food safety with other public concerns. More recently, BWS has been applied to areas such as health (Szeinbach *et al.*, 1999; Flynn *et al.*, 2007), arts (Burke *et al.*, 2007), and cross-country consumer ethics (Auger *et al.*, 2007). Formal proofs of the theoretical attributes of BWS are provided in Marley and Louviere (2005).

Survey

The participants in the study were students in an undergraduate course at the author's employing university in the middle of 2006. The course, for which the author was the lecturer, had an enrolment of 38 students of which 26 attended the class in which the (unannounced) end-of-semester teaching evaluations were executed.

After completion of the University's standard teaching evaluation forms, students were asked to participate in a survey using the Best-Worst questionnaire, presented as "an alternative instrument to the one with which you are familiar". In accordance with University policy, the evaluation exercise was overseen by an administrative support person with the lecturer absent from the classroom.

The nine teaching criteria and a tenth overall evaluation assessment (the latter referred to by Sproule (2000) as the "single-most-important question") contained in the university's standard teaching evaluation form are shown in Figure 1.

	Strongly Agree	Agree	Disagree	Strongly Disagree
1) This lecturer communicated effectively with students (e.g. he/she explained things clearly).				
2) This lecturer stimulated my interest in the subject matter he/she was teaching.				
3) This lecturer encouraged me to think critically.				
4) This lecturer provided helpful feedback to help me learn.				
5) This lecturer encouraged student input and participation during classes.				
6) This lecturer was generally helpful to students.				
7) This lecturer was well prepared and structured the subject matter well.				
8) This lecturer helped me to improve my understanding and/or skills in this subject area.				
9) This lecturer communicated his/her enthusiasm for the subject area.				
10) Overall, I was satisfied with the quality of this lecturer's teaching.				

Figure 1: Questions in standard university evaluation instrument

Figure 2 displays part of the first page of the BWS questionnaire. Students were instructed to select, in each question, the one attribute which applied most to the lecturer and the one attribute that applied least to the lecturer.

The experimental design underlying the BWS questionnaire was based on a Balanced Incomplete Block Design with 12 sets (questions) of three lecturer attributes each. In each question, the three attributes were a different subset of the first nine attributes of the University's standard evaluation form. (Students were requested to provide their rating for the nine teaching criteria only, not the tenth overall teaching rating.) In the experimental design, each attribute co-occurred once with each other attribute (that is, all 36 possible combinations of two attributes appeared once), and each attribute appeared in four subsets. Within each of the 12 sets, order bias was accounted for by ensuring that each attribute appeared approximately equally in each position across the 12 questions.

In the following 12 questions, you are given different combinations of three lecturer attributes. For each of those questions, choose the **ONE** attribute out of those three which applied most to the lecturer of this course, and choose the **ONE** attribute out of those three which applied least to the lecturer of this course. So tick only **ONE** attribute in the 'Most' column and only ONE attribute in the 'Least' column in each question.

Question 1

Most	Attribute	Least
	This lecturer stimulated my interest in the subject matter he/she was teaching.	
	This lecturer helped me to improve my understanding and/or skills in this subject area.	
	This lecturer provided helpful feedback to help me learn.	

Question 2

Most	Attribute	Least
	This lecturer communicated effectively with students (e.g. he/she explained things clearly).	
	This lecturer encouraged student input and participation during classes.	
	This lecturer provided helpful feedback to help me learn.	

Question 3

Most	Attribute	Least
	This lecturer provided helpful feedback to help me learn.	
	This lecturer was well prepared and structured the subject matter well.	
	This lecturer communicated his/her enthusiasm for the subject area.	

Figure 2: BWS instrument – instructions and questionnaire extract

Results and discussion

The sample's student characteristics are shown in Table 1 with the percentages referring to the proportion of the completed number of responses in the total sample size of 26. Clearly, most students in the sample were domestic, male students studying full-time, which is representative of the overall student population in the course.

Table 1: Sample characteristics

Student characteristic	n	%
Mode of study		
Full-time	25	96
Part-time	0	0
Gender		
Female	7	27
Male	18	69
Origin		
Local	23	88
International	1	4

A number of different BWS data analysis methods were carried out and their results compared with each other and with the conventional teaching evaluation results. All measures, including the ranks for the nine teaching criteria in each case, are contained in Table 2 and are now discussed in turn.

Table 2: Evaluation measures (ranks in brackets)

	Mean rating	%(SA+A)	BWS: Most minus Least	BWS: Weights measure	BWS: SQRT (Most/Least)	BWS: WLS
1) This lecturer communicated effectively with students (e.g. he/she explained things clearly).	3.7 (5)	100 (1)	0.885 (4)	10.462 (4)	1.509 (3)	-0.022 (4)
2) This lecturer stimulated my interest in the subject matter he/she was teaching.	3.5 (9)	96 (7)	-1.923 (9)	6.692 (7)	0.492 (7)	-0.950 (9)
3) This lecturer encouraged me to think critically.	3.6 (7)	96 (7)	-1.846 (7)	6.692 (7)	0.475 (8)	-0.874 (8)
4) This lecturer provided helpful feedback to help me learn.	3.6 (7)	96 (7)	-1.846 (7)	6.462 (9)	0.378 (9)	-0.850 (7)
5) This lecturer encouraged student input and participation during classes.	3.8 (2)	100 (1)	0.692 (5)	10.154 (5)	1.237 (5)	-0.024 (5)
6) This lecturer was generally helpful to students.	3.7 (5)	100 (1)	0.192 (6)	9.538 (6)	1.080 (6)	-0.270 (6)
7) This lecturer was well prepared and structured the subject matter well.	3.8 (2)	100 (1)	1.692 (1)	11.808 (2)	2.236 (1)	0.217 (1)
8) This lecturer helped me to improve my understanding and/or skills in this subject area.	3.8 (2)	100 (1)	0.923 (3)	10.731 (3)	1.430 (4)	<i>Base</i> (3)
9) This lecturer communicated his/her enthusiasm for the subject area.	3.9 (1)	100 (1)	1.692 (1)	11.885 (1)	2.094 (2)	0.190 (2)

The second and third columns in Table 2 provide the standard summary measures derived from the conventional teaching evaluation instrument: ‘Mean ratings’ and ‘%(Strongly Agree + Agree)’. Neither the mean ratings measure nor the ‘%(SA+A)’ measure appears to discriminate well between teaching criteria with ranges of 0.4 (on a scale between 0 and 4) and 4 (on a scale between 0 and 100), respectively. There is a clear tendency in the responses towards ‘Strongly Agree’ for all items, well above the midpoint of the rating scale, although it is not apparent whether this tendency is a reflection of extreme response bias towards the affirmative responses or a genuine reflection of student perceptions.

The first of the BWS measures listed in Table 2, ‘Most minus Least’, is an interval-level measure derived from Finn and Louviere (1992). Each student’s rating for each teaching criterion is simply calculated by taking the number of times the student indicated this criterion to be most applicable and subtracting the number of times that the student indicated it as the least applicable across the four subsets in which that criterion occurred. Using the tallies for each student (ranging from -4 to +4 for each criterion), the sample averages were calculated which are shown in the fourth column of Table 2. The figures clearly distinguish the positive from the negative criteria in a relative sense on a scale from -4 to +4. Criteria seven and nine have the highest positive score while criteria two, three and four received the highest negative scores.

The second BWS measure is based on Louviere *et al.* (2004), a study of a choice experiment in which the estimation of choices at the level of individuals is investigated. In essence, this

involves the use of BWS to obtain more information from choice experiments by calculating implied rank weights from respondents' (full) ranking of choice options. From this work, it follows that if there are J options in a set, the implied number of times that the k^{th} ranked option is ranked first among all possible subsets of options (including the single element subset) equals 2^{J-k} . For instance, if there are three options in a set (a , b and c) and option b is indicated as 'best' of the three and option a is indicated as 'worst', there are four implied sets in which b is ranked first ($[b > c > a]$, $[b > a]$, $[b > c]$ and $[b]$), two implied sets in which c is ranked first ($[c > a]$ and $[c]$), and one implied set in which a is ranked first ($[a]$).

After assigning the above rank weights to the survey responses, students' totals for each criterion were averaged over the sample. These averages are displayed in the 'Weights measure' column in Table 2 on a scale from 4 to 16. The relative scoring pattern of the teaching criteria is similar to that of the 'Most minus Least' measure.

The next BWS measure is a ratio scale as discussed in Marley and Louviere (2005). It is derived by, first, dividing the total frequency count of Most by the total frequency count of Least and, then, taking the square root of that ratio. Marley and Louviere (2005) show that this ratio is a ratio scale of the Most frequencies. An important feature of this ratio scale is that it gives a true relative score for each item with a lower boundary score of zero. From Table 2 it can be seen, for instance, that criterion one (effective communication) scored approximately three times as high as criterion two (stimulating interest) and approximately one and a half times as high as criterion six (generally helpful). The overall scoring pattern appears, once again, to be consistent with that of the previous measures.

The final BWS measure is derived from a statistically more rigorous analysis. Following Flynn *et al.* (2007), a sample level paired model of analysis was used. This involves an analysis of all frequency counts of all possible pairs of [Most, Least] combinations. The total number of all possible pairs is $J(J-1)$ which, in the current case, equals 72. A model was estimated using Weighted Least Squares with the natural log of the frequency count of each unique pair as the dependent variable. The natural log of each pair's frequency count is a linear function of the difference in utility (McIntosh & Louviere, 2002). The frequency count of pairs with zero frequencies was adjusted by adding the inverse of the sample size in order for natural logs to be calculated. The estimated parameters are shown in the last column of Table 2. Criterion eight is the base criterion in the estimation. Of the eight parameters that can be estimated from the model, criteria seven and nine have a utility that exceeds that of the base criterion with criteria two, three and four having the lowest negative utility compared with base criterion eight.

All measurement values in Table 2 were compared by way of correlation analysis and the resulting Pearson correlation coefficients are shown in Table 3. Only eight criteria values were estimated in the WLS method; those estimates were compared with the corresponding values on the ratio scale only.

The results show a very strong correlation between all measures. In particular, the correlation between the values of the various BWS variants is high. The high correlation between 'Most minus Least' and 'SQRT(Most/Least)' is consistent with Marley and Louviere (2005) who show that, in comparison with ratio scales, interval-level scales are biased but only to a small extent. A clear implication of the findings is that the computationally straightforward 'Most minus Least' and 'SQRT(Most/Least)' measures can confidently be used as representative

BWS metrics. This has major practical advantages for the analysis of teaching evaluation responses.

Table 3: Correlation between evaluation measures

	Mean rating	%(SA+A)	BWS: Most minus Least	BWS: Weights measure	BWS: SQRT (Most/Least)	BWS: WLS
Mean rating	1					
%(SA+A)	0.853 [*]	1				
BWS: Most minus Least	0.921 [*]	0.951 [*]	1			
BWS: Weights measure	0.919 [*]	0.942 [*]	0.998 [*]	1		
BWS: SQRT (Most/Least)	0.870 [*]	0.840 [*]	0.961 [*]	0.968 [*]	1	
BWS: WLS					0.952 [*]	1

* p<0.01

Table 3 also reveals the high correlation between, on the one hand, the two conventional evaluation measures and, on the other hand, the BWS measures. This is encouraging and reassuring with a view to the long history of conventional teaching evaluation instruments; it suggests that the conventional measures are good indicators of the relative scores of the teaching criteria. However, because there are such small differences between the conventional criteria scores across the teaching criteria reported in this study, it would be tempting to conclude that students' views are about the same for each of the criteria. Yet, the BWS method, by discriminating more clearly between the criteria, highlights that small differences in conventional summary scores hide important differences in student perceptions. In other words, the BWS method acts as a magnifying glass held over the conventional scores. One example is visualised in Figure 3. The conventional mean ratings for criteria six and seven appear to be very similar (3.7 and 3.8, respectively) and may be interpreted as such, even more so since the total percentage of the two 'agree' categories is 100 for both. However, the true relative score, reflected in the BWS ratio score, reveals that the teacher's perceived performance on criterion seven was more than twice the performance on criterion six. This illustrates how the BWS method better highlights the differences among the teaching evaluation items and, hence, enhances the educational evaluation feedback mechanism. The educator can use the BWS results to further explore the underlying issues by way of qualitative methods.

At this point, it is instructive to address briefly the students' feedback to the BWS instrument obtained after the survey. Students generally perceived the BWS questions to be difficult. The main reason given was that, in many questions, two or three teaching criteria were perceived as equally important which made it difficult to choose 'Most' or 'Least'. An alternative explanation is that students are very familiar with the conventional evaluation questionnaire, and complete it routinely, while this was their first exposure to the BWS instrument and, therefore, adopted a more deliberative response approach. At any rate, given the experimental design underlying the BWS survey, if certain teaching criteria were genuinely perceived to be of equal importance, they would be chosen an equal number of times across the questions. However, that was not borne out by the BWS results, which demonstrates that some criteria were actually "more equal than others". While the time length of questionnaire completion was not recorded or requested, students indicated that it took them longer to complete the

BWS questionnaire compared with the standard evaluation instrument. The latter is consistent with the findings in Chrzan and Golovashkina (2006).

	Mean rating	%(SA+A)	BWS: Most minus Least	BWS: Weights measure	BWS: SQRT (Most/Least)	BWS: WLS
1) This lecturer communicated effectively with students (e.g. he/she explained things clearly).	3.7	100.0	0.885	10.462	1.509	-0.022
2) This lecturer stimulated my interest in the subject matter he/she was teaching.	3.5	96.0	-1.923	6.692	0.492	-0.950
3) This lecturer encouraged me to think critically.	3.6	96.0	-1.846	6.692	0.475	-0.874
4) This lecturer provided helpful feedback to help me learn.	3.6	96.0	-1.846	6.462	0.378	-0.850
5) This lecturer encouraged student input and participation during classes.		100.0	0.692	10.154	1.237	-0.024
6) This lecturer was generally helpful to students.		100.0	0.192	9.538	1.080	-0.270
7) This lecturer was well prepared and structured the subject matter well.		100.0	1.692	11.808	2.236	0.217
8) This lecturer helped me to improve my understanding and/or skills in this subject area.		100.0	0.923	10.731	1.430	n/a
9) This lecturer communicated his/her enthusiasm for the subject area.	3.9	100.0	1.692	11.885	2.094	0.190

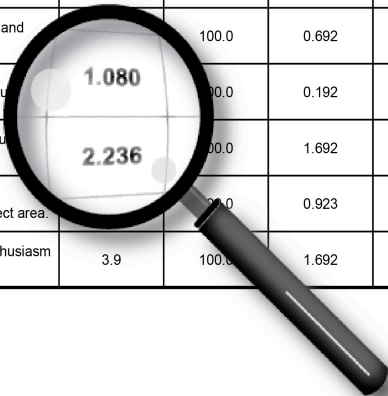


Figure 3: BWS and the magnifying glass effect

Students also commented that the BWS exercise, by forcing them to scrutinise the teaching criteria, highlighted the ambiguity of some of the questionnaire items. For instance, it became apparent that students did not really understand the meaning of the criterion ‘The lecturer helped me to improve my understanding and/or skills in this subject area’? This ambiguity did not become apparent from the conventional evaluation instrument. This is directly related to the issue of the suitability of questions contained in evaluation surveys to capture appropriately the concepts of course quality and/or teacher effectiveness and performance (Boex, 2000; Sproule, 2002; Keeley et al., 2006; Pounder, 2007).

Conclusions

In its pure form, student evaluation of teaching is aimed at providing feedback to educators about the quality and effectiveness of various aspects of their teaching. This formative function of teaching evaluation informs an educator’s reflection about their teaching and areas where their teaching can be improved. For that purpose, having a reliable set of evaluation results in terms of the relative performance levels of various teaching criteria is essential. Conventional rating scales used in teaching evaluation instruments have shortcomings that have the potential to limit their ability to be a guide to teaching appraisal and improvement. In this paper, BWS is proposed as an alternative method that overcomes these shortcomings and more clearly highlights the true, underlying differences in measurements across various teaching evaluation criteria.

The pilot study reported in this paper has produced two main findings. First, there is a high correlation between the conventional rating measures and the BWS metric variants. The implication of this finding is reassuring in that the many conventional teaching evaluations

that have been carried out can, in principle, produce valid measurements. However, the results from such standard evaluation methods need to be interpreted very carefully. In particular, if there is only a small range in the teaching assessment criteria scores, an educator cannot simply assume that they are performing equally well (or badly) across all criteria. A difference of, say, 0.1 on a scale of four should not be dismissed as insignificant. Instead, the differences should be scrutinised carefully to produce insight into the relative performance across the various criteria of interest. The problem is that the conventional measures themselves do not provide any guidance about the true extent of the differences. That is where the discriminating feature of BWS scores comes into play. BWS provides enhanced insights into the genuine relative importance of the teaching criteria. The BWS scores act as the magnifying glass held over the conventional scores and provide a more rigorous indication of relative criteria importances. Of course, the above issues would be of no concern if the results derived from a particular conventional evaluation exercise showed up large, obvious differences between teaching criteria. However, the measures derived from standard evaluation instruments using Likert type scales often do not produce such clear differences.

The second of the study's main findings is that there is a high correlation within the various BWS measures. This implies that there is no need for complicated analyses of the BWS response data. Indeed, simple computations underlying the 'Most minus Least' method or, preferably, the BWS ratio scale, are sufficient to derive robust results. This is very useful for teaching evaluation practitioners, whether they are university administrators carrying out large numbers of teaching evaluations or individual educators carrying out their own evaluation investigations.

To the author's knowledge, the study detailed in this paper is the first application of BWS to student evaluation of teaching so the study should be regarded a pilot study. As such, it is acknowledged that the associated small scale of the survey is a limitation to the findings reported here. It would be of great benefit to replicate this study at a larger scale and in other contexts. One further aspect that could be investigated is the application of BWS to the evaluation of *courses* as opposed to the lecturer's *teaching* of courses. Students may feel less "emotionally attached" to the course as opposed to the lecturer with whom they have had a semester's contact. This is related to the contention about the effect of an educator's relationship with students on students' evaluation of teaching due to, for instance, teachers' characteristics and personality and course grades, (see, for instance, Clayson, 1999; Langbein, 2008). It would be interesting to see if that aspect has any effect on the comparative results of conventional and BWS evaluation measures. In addition to this avenue for future research, it would be interesting to extend the research from an investigation of *performance* criteria to student perceptions of the *importance* of various teaching criteria to teaching effectiveness. This would shine a light on the suitability and nature of the items incorporated in a teaching evaluation instrument. The author is currently pursuing these proposed lines of investigation.

References

- Auger, P., Devinney, T., & Louviere, J. (2007). Using best-worst scaling methodology to investigate consumer ethical beliefs across countries. *Journal of Business Ethics*, 70, 299-326.
- Badri, M., Abdulla, M., Kamali, M., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management*, 20 (1), 43-59.
- Baumgartner, H., & Steenkamp, J. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research*, 38 (2), 143-156.

- Burke, P., Burton, C., Huybers, T., Louviere, J., & Wise, C. (2007). Museum Visitors Care about Everything! Using Best-Worst Scaling for Strategic Focus. *2007 Conference of the Australia New Zealand Marketing Academy*, Dunedin.
- Chrzan, K., & Golovashkina, N. (2006). An empirical test of six stated importance measures. *International Journal of Market Research*, 48 (6), 717-740.
- Clayson, D. (1999). Students' evaluation of teaching effectiveness: some implications of stability. *Journal of Marketing Education*, 21, 68-75.
- Clayson, D., & Sheffet, M. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education*, 28 (2), 149-160.
- Cohen, S. (2003). Maximum difference scaling: improved measures of importance and preference for segmentation. *Proceedings of the 2003 Sawtooth Software Conference*, San Diego, CA, 1-18.
- Cronbach, L. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, 10, 3-31.
- Davies, M., Hirschberg, J., Lye, J., & Johnston, C. (2007, September 12). Let's evaluate the possibilities. *The Australian newspaper*.
- Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education*, 24 (3), 331-350.
- Dolnicar, S., & Gruen, B. (2007). Cross-cultural differences in survey response patterns. *International Marketing Review*, 24 (2), 127-143.
- Finn, A., & Louviere, J. (1992). Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing*, 11 (1), 12-25.
- Flynn, T., Louviere, J., Peters, T., & Coast, J. (2007). Best-worst scaling: what it can do for health care research and how to do it. *Journal of Health Economics*, 26, 171-189.
- Greenwald, A. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, 52 (11), 1182-1186.
- Haskell, R. (1997). Academic freedom, promotion, reappointment, tenure, and the administrative use of student evaluation of faculty (SEF): (Part IV) Analysis and implications of views from the court in relation to academic freedom, standards, and quality of instruction. *Education Policy Analysis Archives*, 5 (21), November 25 1997.
- Keeley, J., Smith, D., & Buskist, W. (2006). The teacher behaviors checklist: factor analysis of its utility for evaluating teaching. *Teaching of Psychology*, 33 (2), 84-91.
- Koh, H., & Tan, T. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management*, 11 (4), 170-178.
- Langbein, L. (2008). Management by results: student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review*, 27, 417-428.
- Louviere, J., Burgess, L., Street, D., & Marley, A. (2004). Modelling the choices of single individuals by combining efficient choice experiment designs with extra preference information. *Working Paper No. 04-005, Centre for the Study of Choice*, University of Technology, Sydney.
- Marley, A., & Louviere, J. (2005). Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology*, 49, 464-480.
- Marsh, H., & Roche, L. (1997). Making students' evaluations of teaching effectiveness effective: The central issues of validity, bias, and utility. *American Psychologist*, 52 (11), 1187-97.
- McDonald, J., & Mills, A. (2007, September 12). Flaws in the SET system. *The Australian newspaper*.
- McIntosh, E., & Louviere, J. (2002). Separating weight and scale value: an exploration of best-attribute scaling in health economics. *Health Economists' Study Group Meeting*, Brunei University, January 2002.
- Onwuegbuzie, A., & Weems, G. (2004). Response categories on rating scales – characteristics of item respondents who frequently utilize midpoint. *Research in the Schools*, 11 (1), 50-59.
- Pounder, J. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education*, 15 (2), 178-191.
- Szeinbach, S., Barnes, J., McGhan, W., Murawski, M., & Corey, R. (1999). Using conjoint analysis to evaluate health state preferences. *Drug Information Journal*, 33, 849-858.
- Sproule, R. (2000). Student evaluation of teaching: a methodological critique of conventional practices. *Education Policy Analysis Archives*, 8 (50), November 2 2000.
- Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review*, 21, 287-294.
- Webster, C. (1990). Evaluation of marketing professors: a comparison of student, peer and self evaluations. *Journal of Marketing Education*, 12 (1), 11-17.

Copyright © 2009 Twan Huybers: The author assigns to HERDSA and educational non-profit institutions a non-exclusive license to use this document for personal use and in courses of instruction provided that the article is

used in full and this copyright statement is reproduced. The author also grants a non-exclusive license to HERDSA to publish this document in full on the World Wide Web (prime site and mirrors) on CD and in printed form within the HERDSA 2009 conference proceedings. Any other usage is prohibited without the express permission of the author.