Author's response to reviews

Title: Estimating preferences for a dermatology consultation using Best-Worst Scaling: comparison of various methods of analysis

Authors:

   Terry Flynn (terry.flynn@bristol.ac.uk)
   Jordan J Louviere (jordan.louviere@uts.edu.au)
   Tim J Peters (tim.peters@bristol.ac.uk)
   Joanna Coast (j.coast@bham.ac.uk)

Version: 2 Date: 28 February 2008

Author's response to reviews: see over
Dear Sirs,

Estimating preferences for a dermatology consultation using Best-Worst Scaling: Comparison of various methods of analysis

Thank you for considering the above paper and for forwarding the helpful comments of the referees. Please find attached a revised version which takes account of these comments. We set out below our responses to each of the referees’ points (which are in bold text). The description of revisions made in response to these are in plain text, with quoted passages from the paper in italics.

Referee #1

The use of so many acronyms in this paper makes it so hard to read. I would argue that only internationally accepted acronyms such as UK and NHS be used and that no methodological acronyms (DCE, WLS) or substantive acronyms (GPSI).

Done.

I know that many use the term discrete choice experiments, but I think that the authors should also use the more general term conjoint analysis. For me these are both forms of conjoint analysis.

By using the term “discrete choice” we have followed terminology in the seminal text by Louviere, Hensher and Swait but following the referee’s suggestion we have added “also called (choice-based) conjoint analysis by many practitioners in North America” to the text on page 4.
The paper is very verbose for applying a simple comparison of two methods (especially the abstract). Cutting back about 30-40% of the text would make the point of the paper so much more apparent.

Given the comments of the second referee and the apparent need not to (simply) refer the reader to the methodological paper for all aspects, we do not feel it would be helpful to make major cuts to the text. We would also draw the referee’s attention to the fact that covariate-adjusted results are presented, in addition to the work comparing methods.

Minor Essential Revisions
The tables are poorly organised, one really needs only the coefficient and something to so with the standard error (the standard error itself, the t-statistic, the P-value or the confidence interval). Using all looks like the authors have just cut and pasted everything from their output.

We now omit the T-ratio and prob value from the output, leaving the coefficient, standard error and 95% confidence interval.

What types of standard errors are being reported? Would it be suitable to make Huber-White corrections?

We thank the referee for this suggestion. Version 9 was the first version of Stata to support robust standard errors for conditional logit and we were unaware of this when submitting the manuscript. Therefore we have repeated the analyses and now report cluster-adjusted (Huber-White) standard errors. This made certain covariates non-significant in the analysis of preference heterogeneity but did not change the main findings.

We have therefore re-written the paragraph beginning “An assumption underlying the regression model results was…” on (what was) page 17 to read: “The Huber-White (robust/sandwich) adjustment inflates the standard errors by the correlation in the error terms (resulting from multiple responses from individuals). Thus, it does not explicitly model heterogeneity in responses. However, the most popular approaches to modelling respondent-specific effects (random effects models) potentially are misleading. Not only do these models ignore other factors that might lead to variation in choice behaviour, virtually all published applications implicitly assumed constant error variances within and between individual responses. If the error variances do vary then the distribution of model estimates is confounded with the distribution of error variances, and the bias can be very large [25]. So, a random effects model represents at best a partial solution, and evidence suggests that this simplistic treatment of heterogeneity is not supported empirically [26]. Furthermore, unlike the case for clinical outcomes, there are no theoretically compelling reasons to support unimodal distributional assumptions for modelling preference heterogeneity.”
We have also amended the first sentence of the paragraph in the discussion beginning “In addressing the limitation of assuming a constant variance it would seem more logical to exploit the power of best-worst methods…” to read:

“To model preference heterogeneity, it can be argued that it is more logical to exploit the power of best-worst methods to make individual-level inferences (estimate parameters for each individual) than attempt to specify them as random effects. Indeed, best-worst scaling can model individual-level utility functions without statistically questionable distributional assumptions about preferences[11]. Similar work will be performed for these data.”

Referee #2
This paper is of interest to researchers with a specific interest in discrete choice experiments and ways of refining the information that can be extracted by this method of preference elicitation.
The paper is well written and the methods and data appear to be sound. Only a few papers have been written within the field of health economics on the topic of best-worst scaling. This is the strength of the paper - but also the potential weakness, since many of the interested readers (including this one) will have heard of best-worst scaling without having a deep understanding of the underlying justification for the approach, nor the specific wordings applied when including best-worst scaling in a questionnaire. On this premise, the paper lacks more precise and detailed descriptions of specific issues.

Major points:
1. On page 5 the authors write: "The additional information provided by the BWS compared with a traditional choice model allows researchers to separate the effects of different attribute impacts across patient sub-groups from those of different level scale values. Such differences can help policy makers decide whether policies to improve levels of key attributes (for instance reduce the incidence of a given side effect of treatment) or those to increase/decrease the perceived impact of attributes themselves (for instance better education to improve patient understanding of a side effect) are the most desirable or feasible". This - as I see it - is the main justification for the BWS approach: that it provides added information that is potentially policy relevant.
I am -however -not entirely convinced by the arguments that are presented. I can see how we, econometrically, can divide the weighing of attributes into two parts:
the general weighting vis-a-vis other attributes, and
the added/reduced weighting associated with the levels of these attributes. But does this necessarily mean that this is how individuals think??? Can we assume that we can launch a policy intervention which only affects the general perception of attribute importance, and not the relative importance of attribute levels? This is a statement which needs some sort of substantiation in terms of reference to - for example - the psychology literature. If such evidence cannot be found, the authors may want to present the statement more as a hypothesis which needs to be tested in future research.
The referee has highlighted a crucial distinction between attribute impact and attribute importance. This issue was explored in a paper that was recently accepted for publication (Marley, Flynn and Louviere, JMP in press). Attribute importance is the more generalisable concept which is of particular interest to psychologists and gives the greatest insight into “how individuals think”. Attribute impact is the ‘mean of the level scale values’ and, as such, is likely to be less generalisable. Therefore we have rewritten this paragraph to:

- make clearer that the key advantage of BWS – its ability to elicit considerably more preference information than a traditional choice experiment – does not rely on the separation of attribute impact and level scale values.
- stress the difference between importance and impact and note that although BWS estimates (only) the latter, this is useful by virtue of the fact that all level scales (and hence, attribute impacts) are estimated on a common scale.
- explain why the separation of attribute impact and level scales is useful on econometric grounds.

Therefore, it now reads:

“The additional information provided by best-worst scaling compared with a traditional choice model is useful to the researcher in three ways.

1. Most importantly, asking about worst as well as best elicits more information about the respondent’s utility function. This increases the researcher’s ability to characterise heterogeneity in preferences and classify respondents into internally homogeneous groups. In certain circumstances best-worst methods can even be used to estimate individual-level utilities [11].

2. For a given attribute, all of the level scale values are estimated (rather than all but one as in a traditional choice experiment) [3]. This allows the researcher to calculate the mean utility across an attribute’s levels – the attribute impact. Furthermore, all attribute levels (across the entire study) are estimated on a common scale, allowing meaningful comparisons of attribute impacts to be made. However, it should be noted that attribute impact is not the same as attribute importance – a more generalisable concept that has been investigated (largely unsuccessfully) by psychologists for 40 years [12]. Nevertheless, because all the level scale values (and hence all the attribute impacts) are on a common scale, knowledge of an attribute’s overall impact is still useful. It may help policy-makers decide whether policies to improve levels of key attributes (for instance reduce the incidence of a given side effect of treatment) or those to increase/decrease the perceived impact of attributes themselves (for instance better education to improve patient understanding of a side effect) are the most desirable or feasible.

3. When effects coding (rather than a dummy variable approach) is used to estimate the level scale values: the econometric model automatically estimates the statistical significance of both level scales and the overall attribute impact.”
2. All attributes - except "waiting time" have only two levels. In the results section (page 12) the authors conclude: “Doctor expertise is clearly the most highly valued attribute whilst convenience is valued slightly more then thoroughness of care. The result of separating overall attribute impact from level scale values is clear: whilst thoroughness of care is not the most important attribute per se, the two levels are very far apart on the utility scale. In contrast, for convenience of attending there is a difference of $2 \times 2.53 = 5.06$ units between the levels of thoroughness of care but only $2 \times 1.02 = 2.04$ units between the level of convenience.” These results can be verified in table 1. This raises some confusion in my mind. If we are dealing with dichotomous variables where the thoroughness attribute for example has the two attribute levels, thoroughness yes and thoroughness no representing the lack or existence of thoroughness can we then in any meaningful way distinguish between the attribute impact and the impact of level scale values??? I can see the distinction when we are dealing with attributes which have several attribute levels, such as waiting time. The authors need to address this issue in more detail, and explain their distinction intuitively.

We thank the referee for making this point. In conjunction with our exposition to address point one, we hope the following additional paragraph in the discussion clarifies this issue:

“The usefulness of attribute impact as a concept is an empirical issue. Because it is simply the arithmetic mean of the levels of an attribute, the impact for an attribute like thoroughness which is defined by its two ‘extreme’ levels (yes/no) is arguably of little value. However, more generally, attribute impact undoubtedly would be useful to researchers; for example, a non-significant scale value coupled with a very small attribute impact (relative to other attributes) may indicate that an attribute is highly disliked per se and respondents do not perceive any difference between levels. Thus, unlike traditional discrete choice experiments, researchers would know that respondents particularly dislike this attribute per se and that policies that attempt to change patient perceptions of the good/service may be more fruitful.”

3. Related to point 2: it needs to be made clearer exactly what the utility function looks like. An equation should be presented in the methods sections such that the relationship between the attribute impacts coefficients and level scale value coefficients are interpreted correctly by the reader.

Done.

4. As far as I can see the attributes and the associated attribute levels are not presented in the manuscript. Nor is there an example of how the BWS question was phrased. Although the authors refer to other publications for a more detailed descriptions, such basic information should be presented in the manuscript.
Please see new table 1 describing attributes and levels. If BMCMRM accept our article, we shall be required to obtain permission from the publisher of our original empirical study for this to be reproduced. The new figure 1 presents an example best-worst question from the study.

Minor points:
1. On page 4 the authors seek to describe the difference between best worst scaling and standard DCE applications using the following description: "BWS represents respondents with scenarios one at a time. Rather than comparing the utility of entire scenarios, respondents evaluate and compare the utilities of the attribute levels on offer within one scenario, picking that pair of attribute levels that maximise the difference in utility between them.". It is very difficult to follow this description - mainly on the ground of imprecise terminology. What is a scenario exactly, and what is an "entire scenario"?? Would it be useful to distinguish between alternatives and scenarios (which are a set of alternatives)? And what is exactly meant by "a pair of attribute levels"??

We have amended this to read:
"...best-worst scaling presents respondents with profiles (in this case appointments) one at a time. Respondents make choices within profiles (appointments) rather than between profiles. Thus, for a given profile, the set of alternatives on offer comprises the attribute levels that define that particular profile (appointment). By choosing the best and worst attribute levels on offer within that profile, respondents select that pair (best and worst) of attribute levels that lie furthest apart on the latent utility scale. Thus, variations on best-worst scaling have appeared, sometimes called “maximum difference scaling[8].”

2. On page 4 the authors write "Best-worst scaling........is another solution". It is confusing that the authors refer to "another" solution - since it appears to refer to a previous solution having been presented. In reality the authors only seem to be referring to "several methods presented in the literature".

This has been amended to read:
"Best-worst scaling........is a solution that utilizes a different design of choice experiment."

3. Page 7: In each scenario (appointment offered) respondents were asked to choose one attribute that was best and one that was worst, based on the levels described in the scenario. Thus each choice presented a pair of attribute levels. On page 9 on the use of the marginal model conditional logit analysis the authors further explain: The outcome variable is coded equal to one for the chosen best outcome and coded equal to zero for the remaining (non-chosen) attribute levels for a particular choice set and individual. First: does this mean that the information on the worst attribute is redundant??
Second: in the first sentence is referred to scenario (one appointment offered) in the second is referred to a choice set (which must involve two or more possible appointments)??

We were imprecise in our explanation and have amended this to explain how information from the worst choices is used:

“Each chosen attribute level was expanded to the 2K=8 attribute levels (4 best and 4 worst) available in each choice set (profile). The outcome variable is coded equal to one for the chosen outcome (whether a best attribute level or a worst attribute level) and equal to zero for the remaining (non-chosen) attribute levels for a particular profile (appointment) and individual. The independent variables were coded with a sign change for all observations pertinent to the worst choice data, to reflect the reciprocal relationship between best and worst probabilities. It should be noted that any inference using the log-likelihood from the marginal method is potentially misleading, since the likelihood function assumes best and worst choices are made independently.”

We have also amended the second paragraph of the discussion:

“The study also provides guidance about choice of analytic model. It demonstrated that although the marginal method of analysis makes potentially unwarranted assumptions about independence of best and worst choices and the nature of the error terms in the utility model, it provides a good approximation to the paired (maxdiff) model analysis, suggesting that the resulting estimates are consistent.”

Given the apparent confusion over the phrase “choice set”, we have also included an additional figure (see figure 1) giving an example best-worst question from the study. This should help familiarize readers with the nature of the best-worst task.

We trust that the referees’ comments have been dealt with satisfactorily, and look forward to hearing from you in due course.

Yours faithfully,

Joanna Coast PhD
Terry N Flynn PhD
Jordan J Louviere PhD
Tim J Peters PhD