Reviewer's report

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

Version: 1 Date: 23 January 2008

Reviewer: Dorte Gyrd-Hansen

Reviewer's report:

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.
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 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.

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.

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.

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"?

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".

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)??

What next?: Accept after minor essential revisions

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.