WORK IN PROGRESS; PLEASE DO NOT QUOTE WITHOUT AUTHORS' PERMISSION

CONJOINT MEASUREMENT IN HEALTH ECONOMICS: A CONCEPTUAL AND EMPIRICAL CONSIDERATION OF VALIDITY AND RELIABILITY

Stirling Bryan[1]
Rob Sheldon[2]
Martin Buxton[3]

[1] Health Economics Facility
University of Birmingham
Park House
40 Edgabaston Park Road
Birmingham B15 2RT

[2] Accent Marketing & Research
Gable House
14-16 Turnham Green Terrace
London W4 1QP

[3] Health Economics Research Group
Brunel University
Uxbridge
Middlesex UB8 3PH

Paper presented to the Health Economists' Study Group Meeting, University of Birmingham, 6-8 January 1999

1. INTRODUCTION

In health care there exist few opportunities to observe data that clearly reveal individual preferences. This is largely a result of the presence of information asymmetry in the market leading to a situation of decision-making being dominated by clinicians rather than patients. Thus, often the only course of action open to the researcher is to use data on stated preferences rather than observed actions. In seeking measures of preference in hypothetical contexts, one method that can be used is 'conjoint measurement'. Conjoint measurement techniques refer to a number of different approaches all of which use peoples' statements of how they would respond to different hypothetical situations. The term 'conjoint analysis' means the decomposition of individual evaluations of a designed set of multi-attribute alternatives into the separate components of utility (Green & Srinivasan, 1978). In conjoint exercises respondents rate, rank or choose between alternatives. The analyst designs a set of hypothetical alternatives based on a limited set of 'important' attributes, and obtains from the respondent an indication of the relative preference for each alternative. The simplest indicator of preference involves the selection of one alternative from two options, and the exercise is repeated with the values of the attributes within the alternatives being systematically altered.

Conjoint methods were originally developed in marketing research in the early 1970s and have been widely used, especially in the transport research field (Kroes & Sheldon, 1988; Hensher et al, 1988, Magat et al, 1988; Hensher, 1994). In health economics research there have been few applications of conjoint techniques until recently (Ryan & Hughes, 1997; Bryan et al, 1998; Vick & Scott, 1998).

Measurement reliability and validity are important aspects of preference elicitation where conjoint methods, or any other approach, is used. The next section of this paper presents a framework for the consideration of reliability and validity issues pertinent to conjoint measurement when applied in the health care sector. This is followed by a discussion of issues of measurement of reliability and validity, and finally, an empirical examination of one aspect of conjoint reliability is then reported.

2. CONCEPTS OF RELIABILITY AND VALIDITY IN CONJOINT MEASUREMENT

2.1 Variation in conjoint measurement

In considering conjoint reliability and validity it is helpful to think in terms of the conventional distinction between two possible sources of variance in measurement: systematic and random.

Systematic variance

Systematic sources of variance tend to influence the conjoint measurements in one direction. The error they introduce is continuous, causing the measurements to be systematically biased.

Random variance

Random variance can cause the conjoint measurements to vary in either direction and tends to be self-compensating in aggregate across respondents or in the long-run where several repeat measurements are taken.

Thus, an observed measurement, in a conjoint analysis or any other measurement exercise, includes the following sources of variation (Bateson, Reibstein and Boulding, 1987).

$$X_o = X_t + X_s + X_r$$

where: X_o = observed score,
X_t = true score,
X_s = systematic sources of error, and
X_r = random sources of error.

Validity is then defined as a situation where there is no error in measurement (i.e. $X_s = 0$ and $X_r = 0$) such that

$$X_o = X_t$$

Reliability is defined as the consistency or reproducibility of results such that a reliable measure has no variation in X_o due to random errors (i.e. $X_r = 0$) but may have systematic sources of error (i.e. $X_s \neq 0$). Thus, reliability only requires that

$$X_o = X_t + X_s$$

Therefore, reliability is a necessary but not sufficient condition for the presence of validity.

2.2 Reliability in conjoint measurement

Thus, the reliability of an instrument is concerned with the extent to which it is free from random error and may be considered as the amount of a score that is signal rather than noise. Random variance represents error that can come from a number of sources: "chance variation, fatigue, temporary change in test conditions, fluctuations in memory or preference, and all other factors that are temporary and shifting" (McCullough and Best, 1979). In conjoint measurement the principal focus for the assessment of reliability, or random variance, is 'reproducibility', sometimes referred to as measurement 'stability' (Fitzpatrick *et al*, 1998). Reproducibility assesses the extent to which subjects can replicate their judgements over some specified time interval, where the measurement instrument remains unchanged (i.e. test-retest reliability). The reliability of a particular measurement tool, such as conjoint analysis, is not a fixed property but is dependent upon both the context and the population studied (Streiner and Norman, 1995).

2.3 Validity in conjoint measurement

The validity of a measurement instrument is an assessment of the extent to which it actually measures what it claims to measure. There are a number of different ways of establishing the validity of a measure. As with reliability, validity is not a fixed property (Jenkinson, 1995). It is, therefore, not possible to talk about conjoint analysis as being a valid or invalid measure. Here we use a psychometric categorisation of validity as a framework for considering issues of validity in the use of conjoint measurement in health economics.

2.3.1 *Criterion validity*

Criterion validity is the 'highest level' assessment of validity. It concerns the extent to which there is agreement between the measurement instrument in question, here conjoint analysis, and another measure generally accepted as a more accurate or 'criterion' variable. The assessment of criterion validity is often difficult, since it is rare that a perfect 'gold standard' measure exists (Fitzpatrick *et al*, 1998). Thus, it is more often the case that less direct approaches to validity assessment are adopted, such as investigations of construct, face and content validity. In empirical conjoint measurement work there exist inherent difficulties associated with obtaining a measure of the respondent's true underlying preference (i.e. X_t in the above model). Revealed preference exercises are especially problematic in the health care context because of the presence of information asymmetry in the market.

2.3.2 *Construct validity*

An alternative approach is to evaluate the construct validity of conjoint measures (Gegax and Stanley, 1997). Construct validity concerns the degree to which the preferences measured in conjoint analysis relate to other measures which, on the basis of current understanding, are known to be related to the core 'construct', that is, respondent preferences. Streiner and Norman (1995) indicate that 'the burden of evidence in testing construct validity arises not from a single powerful experiment, but from a series of converging experiments.'

The most robust means by which construct validity can be tested is through the measurement of 'convergent and discriminant validity' (Campbell and Fiske, 1959; Fitzpatrick *et al*, 1998). The assessment of convergent validity requires the investigation of agreement between different measurement instruments that are all intended to measure the same feature. The assessment of discriminant validity requires the investigation of the absence of agreement between instruments which are designed to measure different features.[1] In terms of conjoint measurement, convergent validity is most relevant (Bateson, Reibstein & Boulding, 1987).

A commonly employed definition of convergent validity is the agreement between two attempts to measure the same feature through 'maximally different' methods. In contrast, reliability can then be defined as agreement between two efforts to measure the same feature through 'maximally similar' methods (Campbell & Fiske, 1959). The distinction between reliability and convergent validity, therefore, rests on the appropriate definitions of 'maximally similar' and 'maximally different'. Precise definitions for these terms are difficult and, thus, a spectrum extends from convergent validity to reliability, along which the similarity of methods varies. The investigation of convergent validity provides an opportunity to quantify the magnitude of the systematic error.

2.3.3 *Face and content validity*

The investigation of face and content validity of a measurement tool largely involves qualitative matters of judgement. There is no formal quantitative process by which

[1] Some would use the term 'discriminant validity' to refer to a situation where between group differences that were hypothesised, a priori, are observed empirically. Others refer to this form of validity as 'theoretical validity'.

these aspects of validity can be assessed. Face validity refers to 'what an item appears to measure based on its manifest content' and content validity refers to 'how well a measurement battery covers important parts of the health components to be measured' (Ware *et al*, 1981). These two aspects of validity can only be explored by the examination of the questionnaire. In conjoint analysis, where different questionnaires and measurement instruments tend to be constructed for each study, these aspects of validity can only be commented upon for each individual study separately.

2.4 Issues in the empirical assessment conjoint reliability and validity

The focus for the discussion of empirical issues in this paper is on convergent validity and reliability. Studies that constitute empirical assessments of convergent validity or reliability differ in terms of the tasks being undertaken by the respondents. The following definitions are used here. *Reliability studies* involve two parts (the main task and the reliability task) both adopting the same or very similar methods. *Convergent validity studies* also involve two parts (the main task and the convergent validity task) where the two parts have structural differences between them.

Conjoint analysis reliability and convergent validity studies, thus, require replications of data collection involving the changes in the time of administration *and* or changes in the items selected. Thus, alternative forms of reliability and convergent validity must be considered. Reliability is principally an issue of reliability over time, and there are three principal forms of convergent validity: over attribute set, over stimulus set, and over data collection method (Bateson, Reibstein & Boulding, 1987).

Reliability over time is assessed by repeating a given measure using the same respondents and precisely the same instruments at two separate points in time. Thus, the only aspect varied is the time of administration. An important issue to be addressed is the appropriate length of time between administration of the main and reliability questionnaires. A short interval risks memory effects and so apparent reliability may be misleading, whereas a long interval risks changes in the underlying preferences being measured.

Convergent validity over attribute set assesses the extent to which the preferences for certain attributes depend on the other attributes or levels in the conjoint exercise. This is tested by examining the stability of the preferences for attributes that are common when other attributes in the stimuli are varied. (Bateson *et al*, 1987). This issue can be investigated in a number of ways:

- changing the nature of one of the attributes; or
- varying the number of attributes; or
- varying the number of levels but keeping the attributes the same; or
- varying both the number of attributes and number of levels.

Convergent validity over stimulus set is concerned with the use of fractional factorial designs in conjoint measurement. Such designs are almost always a necessary component in conjoint analysis design but raise additional methodological questions relating to the stability of preferences to the fractional factorial chosen.

Convergent validity over data collection procedure stems from the fact that a wide range of data collection procedures can be used for conjoint analysis. The key methodological question is whether the results of the conjoint measurement exercise are different when different data collection procedures are used.

3. MEASUREMENT ISSUES

This section of the paper considers issues of measurement of reliability and validity. A variety of measurement methods are available and it is helpful to make a distinction between measurement at the *input data level* and at the *results level.*

3.1 Measurement of reliability / convergent validity at the input data level

The most straightforward approaches to reliability and validity measurement are at the input data level. Many studies have measured the correlation between the raw data from the two administrations of a conjoint measurement instrument (Segal, 1982; Leigh *et al*, 1984). It is increasingly accepted in the world of medical research that correlation is an inappropriate form of analysis as an indicator of 'good agreement' (Bland & Altman, 1986; Altman, 1991). The correlation coefficient is simply a measure of the strength of linear association between two variables. An alternative measure of agreement at the data input level is the proportion of respondents giving a particular response in the two replications. For example, in the case of the discrete choice method where the choice is always between two scenarios (A and B), comparison can be made between replications in terms of the proportion of respondents choosing A or the proportion choosing B, using either McNemar's or a Chi^2 test as appropriate (Altman, 1991).

The simplest approach to assessing agreement is to see how many exact matches were observed between replications. However, this approach takes no account of where the agreement was, and that some agreement would be expected by chance. Thus, a more meaningful measure considers the agreement in excess of that which would have been expected by chance. The name for this measure of agreement is kappa (K) which has a value of 1.00 when agreement is perfect, a value zero indicates no agreement better than chance. There are no absolute thresholds for K indicating good agreement but the following guidelines tend to be used (Altman, 1991): poor agreement $K<0.20$, very good agreement $K>0.80$.

3.2 Measurement of reliability / convergent validity at the results level

Other measures of conjoint reliability / convergent validity require the estimation of conjoint parameters and involve the comparison at the *results level.* Some alternative approaches are outlined below.

An assessment can be made of the similarity of the coefficients estimated for each attribute in the two aggregate regression models, one estimated using data from the main task and the other using data from the reliability / convergent validity task. This involves the estimation of the two aggregate regression models and the comparison of the two coefficients for each attribute, one from each model. If the confidence intervals around the coefficient estimates overlap then this provides a broad indication of no difference between coefficient estimates.

Another option in the measurement of reliability / convergent validity involves the estimation of two regression models for every respondent: the main task model and the reliability task model. A comparison can then be made of the two vectors of regression coefficients for each attribute. The comparison of the two vectors can be undertaken by estimating a correlation coefficient. However, given that each model will be estimated using only a small number of observations, the variance associated with each coefficient estimate will be large and the correlation is unlikely to provide a stable indicator of reliability. An alternative approach is to calculate the difference between coefficient values within each respondent for each attribute and establish whether the confidence interval around the mean for the vector of differences crosses zero: a mean value significantly different from zero would imply a difference between coefficient vectors.

Finally, the results from the aggregate model, estimated using data from the main task, can be used to predict the responses to the scenarios presented in the reliability / convergent validity task. Once the attribute coefficients from the two tasks have been computed, the coefficients from one task can be used to predict the responses to the choices posed in the other. Predicted and observed scores can then be compared.

4 EMPIRICAL ASSESSMENT OF RELIABILITY

4.1 Methods

4.1.1 Study design

This section of the paper describes an empirical assessment of reliability of conjoint measurement applied in a health care context. Two investigations of test-retest reliability were undertaken: one involved an assessment of reliability immediately after the original responses were provided, and the other involved an assessment of reliability after approximately two weeks.

The sample and methods adopted in this study were similar in many respects to those reported in Bryan *et al* (1998). The exercise was again framed in terms of diagnosis and treatment options for patients with knee injuries. However, this survey was undertaken on a larger sample (n=585), all scenarios were described in terms of 4 attributes and the levels on the attributes had been amended from the earlier study. The attributes and levels used in this study are described in Table 1.

The target population chosen for the study was undergraduate students thought likely to be at risk of a knee injury: undergraduate sports scientists or sports teachers. Data were collected from five UK universities (Brunel, Exeter, Kingston, Staffordshire and Birmingham) using questionnaires completed either at the end or at the beginning of a lecture after a short introductory talk.

Subjects were initially asked to complete a questionnaire in two parts: each part contained 8 choices. The 8 choices in the first part (Questionnaire 1) were all original whereas 4 of the 8 choices in the second part (Questionnaire 2) were exact duplicates of choices that had been presented a few moments earlier. Twelve data points per respondent were used for the baseline analysis which were the responses to the first 12 original choices presented in the two questionnaires.

In an attempt to minimise the chance of bias being introduced, respondents were not told they were subjects of a test-retest reliability exercise and were initially presented with Questionnaire 1 and only once it had been completed were they presented with Questionnaire 2. At the time Questionnaires 1 and 2 were completed, respondents were asked to provide a contact address if they were happy to complete a third, postal, questionnaire (Questionnaire 3). Those respondents who provided their contact details were then sent Questionnaire 3 with a pre-paid return envelope after approximately two weeks. Non-responders were sent a single reminder. The questionnaire contained 12 choices which were exact duplicates of the original 12. The precise nature and extent of the duplication of choices across Questionnaires 1, 2 and 3 is detailed in Table 2.

Respondents to Questionnaire 3 were self-selected and this may have introduced some bias. For example, one might hypothesise that respondents who found the initial questionnaires difficult to complete would be less likely to provide their contact details. In order to assess the extent of such bias, the level of association between providing contact details and failing the initial reliability test was investigated.

If a different pattern of choices was found in Questionnaire 3 then one possible explanation would be a different level of experience of relevant factors between completion of Questionnaires 1 and 2, and the subsequent completion of Questionnaire 3. In order to establish whether levels of experience of relevant factors had changed, respondents to Questionnaire 3 were asked whether they had suffered a knee injury, received an MRI scan or undergone knee surgery since the first survey.

The analysis undertaken of the data collected in this study had four main stages:

- the identification of individuals who responded in different ways to the duplicate choices;
- an assessment of the level of reliability at the input data level;
- an assessment of the importance of the unreliable data at the results level;
- an investigation of the characteristics associated with unreliable respondents.

4.1.2 Analysis: identification of unreliable respondents

The various comparisons made in order to identify unreliable respondents and responses are described in Table 3. The first stage in the analysis was to identify the number of individuals who responded in an unreliable manner.

Comparison 1

For the immediate reliability check comparison was made of the responses to the original 4 choices in Questionnaire 1 (choices Q1C1, Q1C4, Q1C5 and Q1C6) with the responses to the duplicate 4 choices in Questionnaire 2 (choices Q2C2, Q2C4, Q2C6, and Q2C8).

Comparison 2

For those who had responded in a completely reliable manner to the immediate reliability checks, comparison was made of the responses to the original 12 choices (taken from Questionnaires 1 and 2) with the responses to the 12 duplicate choices included in Questionnaire 3.

Comparison 3
For those who had responded in an unreliable manner to the immediate reliability checks (i.e. where at least one of the responses to the 4 duplicate choices in Questionnaire 2 were different from the responses to the original 4 choices in Questionnaire 1), comparison was made of the responses to the 8 choices that appeared only once in Questionnaires 1 and 2 with the responses to those same choices in Questionnaire 3.

4.1.3 Analysis: reliability at the input data level

The level of agreement between respondents at the input data level was assessed using McNemar's test and kappa. Both were performed for the immediate reliability check (Comparison 1) and for the follow-up questionnaire at 2 weeks (Comparison 4).

4.1.4 Analysis: reliability at the results level

Two approaches to analysis of reliability at the results level were used: comparison of regression coefficients from the aggregate models and comparison of regression coefficients for individual respondent models.

Aggregate Models
The first assessment involved the re-estimation of the conjoint analysis probit models using different data-sets. A summary of the alternative data-sets used is provided in Table 4. The modelling approach used here was random effects probit, in line with other recent conjoint analyses (Ryan & Hughes, 1997; Bryan et al, 1998; Vick & Scott, 1998). The same model specification as reported in Bryan *et al* (1998) was used with the choice responses as the binary dependent variable and the difference in levels for each attribute as the independent variables (COST, ARTH, TIME and RES).

The baseline model was estimated using a total of 12 data points per respondent, which were the responses to the first 12 original choices presented in Questionnaires 1 and 2. The importance of the discrepancies in responses were investigated by re-estimating the model using the data from the duplicate questions in Questionnaire 2, in place of the original data from Questionnaire 1. This re-estimated model was defined as Model 1 (see Table 4).

The importance of the discrepancies in responses between Questionnaires 1 and 2 taken together, and Questionnaire 3 were investigated by re-estimating the model again, including only data from Questionnaire 3. Given that only a sub-sample responded to Questionnaire 3, the baseline model was also re-estimated using only data from respondents who returned a completed copy of Questionnaire 3 (Model 2). This allowed the comparison to focus on the importance of reliability rather than being confounded by different samples of respondents.

Individual Respondent Models
The second assessment of reliability at the results level involved the estimation of two regression models for every respondent who completed all three questionnaires. Thus, a total of 251 pairs of regression models were estimated; each pair constituted the model estimated using data from Questionnaires 1 and 2 (data from the first 12 original choices only), and the model estimated using data from Questionnaire 3. Each model gave a separate estimate of each attribute coefficient for all 251

respondents. Therefore, two vectors of regression coefficients were estimated for each of the model attributes and the constant term.

For each attribute and the constant, the vector estimated using data from Questionnaire 3 was subtracted from the vector estimated using data from Questionnaires 1 / 2. This gave a single vector for each attribute and the constant that reflected the difference in coefficient values between data collection rounds. If respondents gave exactly the same responses to all 12 choices in the two data collection rounds then the differences in coefficient values would be zero, for all attributes. Therefore, to assess the importance of variation in the responses provided, confidence intervals were estimated for each vector of differences and where they crossed zero, this was taken as an indication of no statistically significant difference between data collection rounds.

4.1.5 Analysis: identification of the characteristics associated with unreliable respondents

The information on changes in experience of relevant factors was used to predict unreliable respondents. Additionally, other factors such as a respondent's age, sex, level of sporting activity and unreliability in Questionnaire 2 were also considered. This analysis involved the estimation of a probit model where the binary dependent variable was defined by whether a respondent was or was not reliable in their responses in Questionnaire 3, and the independent variables related to changes in experience of knee injuries or treatments between data collection rounds and other respondent characteristics.

4.2 Results

4.2.1 Response rate / Sample

The initial survey (Questionnaires 1 and 2) included 585 respondents. For the follow-up survey (Questionnaire 3), a total of 363 respondents, out of the total sample of 585 (62%), expressed a willingness to help further in the research project and provided their contact details. A significantly higher proportion of respondents who gave their contact details were defined as unreliable in Comparison 1, compared to the proportion that did not give their contact details (McNemar's test, p=0.001). This indicates the potential for some selection bias to have been introduced, although the precise level is difficult to quantify. Of the 363 respondents who gave their details, 251 (69%) returned Questionnaire 3. Thus, the overall response to the third part of the survey was 43% (251/585).

Only a small proportion of respondents to Questionnaire 3 (n=19; 7.6% of all respondents) had a different level of experience at the time they completed that questionnaire. The majority of those with a different level of experience had recently injured their knee (n=14).

4.2.1 Identification of unreliable respondents

Table 5 shows the number of individuals who gave responses to the duplicate choices in Questionnaire 2 that were different to their responses to the same choices in Questionnaire 1 (Comparison 1). The majority of respondents (almost 60%) made choices that were completely reliable between Questionnaires 1 and 2. The proportion who gave different responses for more than one choice was only 13% (n=75) in total.

Table 5 also reports data on reliability at 2 weeks. For those respondents who provided completely reliable responses at the immediate reliability check (Comparison 2), only 30% continued to respond in a completely reliable manner, although almost 65% either replied reliably or responded differently for only one (out of the 12) duplicate choices. For those respondents who failed to respond reliably for the immediate check (Comparison 3), the proportion of respondents who continued to respond in an unreliable manner (approximately 68%) is similar to the proportion of respondents who initially responded in a reliable manner but did not continue to do so (approximately 71%). Thus, from these aggregate data, it does not seem that unreliable respondents in Questionnare 3 were more likely to have been unreliable respondents at the immediate reliability check.

4.2.3 Reliability at the input data level

Table 6 shows the total number of responses across all respondents for the 4 choices that were the same in Questionnaires 1 and 2. A total of 317 choices (13.77%) were answered in a different way in Questionnaire 2 compared to Questionnaire 1. The McNemar's test found there to be no significant difference between proportions for the data from the initial two questionnaires ($p>0.05$). Table 7 shows the total number of responses across all respondents that were different for the 12 choices repeated in Questionnaire 1 / 2 and Questionnaire 3. (The comparison involved only the data from the responses to the first 12 original choices in Questionnaires 1 and 2). A total of 407 choices (13.62%) were answered in a different way in Questionnaire 3 compared to Questionnaires 1 / 2. Whilst the difference between proportions was small it was found to be statistically significant ($p<0.01$).

The number of agreed responses to the 4 duplicate choices in Questionnaires 1 and 2 was 1984 (86%). The number of agreed responses that would have been expected by chance is 1173 (51%). This gives a kappa statistic of 0.71 (i.e. "good agreement"). The number of agreed responses to the 12 duplicate choices in Questionnaires 1 / 2 and Questionnaire 3 was 2581 (86%). The number of agreed responses that would have been expected by chance is 1798 (60%). This gives a kappa statistic of 0.65 (i.e. "good agreement" between replications).

4.2.4 Reliability at the results level

Aggregate Models

Table 8 and Figures 1 to 5 give the results of the conjoint analysis random effects probit models. For all attributes, the use of alternative data-sets with varying levels of reliability, had no marked effect on the coefficient values: there were no significant differences between coefficients estimated using the different models.

Individual Respondent Models

Table 9 reports the results of the comparisons of the two vectors of coefficients from the individual respondent models: one estimated using data from Questionnaires 1 and 2 and the other using data from Questionnaire 3. The results indicate that for all model attributes and the constant term, there are no statistically significant differences: the 95% confidence intervals for the differences between means coefficient values for all variables cross zero. Thus, the data collected using Questionnaire 3 appears to provide a range of individual respondent models that are similar in important respects to those obtained using data collected at the initial round.

4.2.5 Identification of the characteristics associated with unreliable respondents

A total of 19 respondents to Questionnaire 3 indicated that they had recently (i.e. since they completed the first part of the survey) had one of the following: a knee injury, an MRI or an arthroscopy. Of the 19, only 6 were found to give responses that were different in Questionnaire 3 compared to their responses to Questionnaires 1 and 2. Thus, it would appear that changing levels of experience of the factors being investigated was not a strong influence on the reliability of the responses provided in the postal questionnaire.

The probit model that was estimated to investigate which factors were associated with unreliable respondents is reported in Table 10. Only two of the factors included in the model had a coefficient that was significantly different from zero (at the 0.05 significance level): recent experience of arthroscopy and the reliability of responses in Comparison 1. Recent knee surgery increased the probability of the respondent giving different responses at follow-up, as would be expected. However, those respondents who were reliable in their responses in Comparison 1 were significantly more likely to respond in an unreliable manner in the follow-up survey. This is a counter-intuitive finding that is difficult to understand. The authors would welcome comments on this puzzling result.

5 SUGGESTED POINTS FOR DISCUSSION AT HESG

1. Is a focus on convergent validity and reliability in conjoint work appropriate?

2. Are there alternative/supplementary approaches to the measurement of convergent validity and reliability?

3. Comments on the empirical assessment of reliability reported here:
    - robustness of results?
    - generalisability of results?
    - explanations for factors associated with reliable respondents?

REFERENCES

Altman DG (1991) Practical Statistics for Medical Research, London: Chapman & Hall

Bateson JE, Reibstein D, Boulding W (1987) Conjoint Analysis Reliability and Validity: a framework for future research. In MJ Houston *Review of Marketing*, Chicago: American Marketing Association

Bland JM, Altman DG (1986) Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement, *Lancet*, February 8, pp307-310

Bryan S, Buxton M, Sheldon R, Grant A (1998) Magnetic Resonance Imaging for the Investigation of Knee Injuries: An Investigation of Preferences, *Health Economics*, vol7, pp595-603

Campbell DR, Fiske DW (1959) Convergent and Discriminant Validation by the Multitrait - Multimethod Matrix, *Psychological Bulletin*, vol56, pp81-105

Fitzpatrick R, Davey C, Buxton MJ, Jones DR (1998) Evaluating patient-based outcome measures for use in clinical trials, *Health Technology Assessment*, vol12, no14

Gegax D, Stanley RL (1997) Validating Conjoint and Hedonic Preference Measures: Evidence from Valuing Reductions in Risk, *Quarterly Journal of Business Economics*, vol36, no2, pp30-55

Green PE, Srinivasan V (1978) Conjoint Analysis in Consumer Research: Issues and Outlook, *Journal of Consumer Research*, vol5, pp103-212

Hensher DA (1994) Stated preference analysis of travel choices: the state of practice. *Transportation* vol21, pp107-133.

Hensher DA, Barnard PO, Truong TP (1988) The Role of Stated Preference Methods in Studies of Travel Choice, *Journal of Consumer Research*, vol5, pp45-58

Jenkinson C (1995) Evaluating the efficacy of medical treatment: possibilities and limitations, *Social Science and Medicine*, vol41, pp1395-1401

Kroes EP, & Sheldon R (1988) Stated Preference Methods: An Introduction, *Journal of Transport Economics and Policy*, vol22, no1, pp11-26

Leigh TW, MacKay DB, Summers JO (1984) Reliability and Validity of Conjoint Analysis and Self Explicated Weights: A Comparison, *Journal of Marketing Research*, vol21, pp456-462

Magat WA, Viscusi WK, Huber J (1988) Paired comparison and contingent valuation approaches to morbidity risk valuation, *Journal of Environmental Economics and Management* vol15, pp395-411

McCullough JL, Best R (1979) Conjoint Measurement: Temporal Stability and Structural Reliability, *Journal of Marketing Research*, vol16, pp26-32

Peter JP (1979) Reliability: A Review of Psychometric Basics and Recent Marketing Practices, *Journal of Marketing Research*, vol16, pp6-17

Reibstein D, Bateson JEG, Boulding W (1988) Conjoint Analysis Reliability: Empirical Findings, *Marketing Science* vol7, no3, pp271-286

Ryan M, Hughes J (1997) Using conjoint analysis to assess women's preferences for miscarriage management, Health Economics vol6, pp261-273

Segal MV (1982) Reliability of Conjoint Analysis: Contrasting Data Collection Procedures, *Journal of Marketing Research*, vol19, pp139-143

Streiner DL, Norman GR (1995) Health measurement scales: a practical guide to their development and use (2nd Edition), Oxford: OUP

Vick S, Scott A (1998) Agency in health care: Examining patients' preferences for attributes of the doctor--patient relationship, *Journal of Health Economics*, vol17, no5, pp587-605

Ware J, Brook RH, Davies AR, Lohr KN (1981) Choosing measures of health status for individuals in general populations, *American Journal of Public Health*, vol71, pp620-625

Wittink DR, Reibstein D, Boulding W, Bateson JE, Walsh JW (1989) Conjoint Reliability Measures, *Marketing Science*, vol8, No4, pp371-374

Table 1 Attributes and levels used in main study

Attributes	*Levels*
Treatment	100% chance of requiring an arthroscopy 80% chance of requiring an arthroscopy 60% chance of requiring an arthroscopy
Time from initial consultation to end of treatment process	6 weeks 12 weeks 18 weeks 24 weeks
Resolution of knee problem	90% chance that knee problem is completely resolved 80% chance that knee problem is completely resolved 70% chance that knee problem is completely resolved 60% chance that knee problem is completely resolved
Total cost of MRI to the patient	zero £200 £400 £600

Table 2 Duplication of choices across Questionnaires 1, 2 and 3

Choice	*Questionnaire 1*	*Questionnaire 2*	*Questionnaire 3*
1	Q1C1	Q2C6	Q3C1
2	Q1C2	not included	Q3C2
3	Q1C3	not included	Q3C3
4	Q1C4	Q2C4	Q3C4
5	Q1C5	Q2C2	Q3C5
6	Q1C6	Q2C8	Q3C6
7	Q1C7	not included	Q3C7
8	Q1C8	not included	Q3C8
9	not included	Q2C1	Q3C9
10	not included	Q2C3	Q3C10
11	not included	Q2C5	Q3C11
12	not included	Q2C7	Q3C12

NB Each row in the table relates to the same choice. Thus, Q1C1 refers to choice number 1 presented in Questionnaire 1 which was duplicated in as choice 6 in Questionnaire 2 and as choice 1 in Questionnaire 3.

Table 3 **Identification of unreliable respondents and responses: comparisons of data**

Comparison	*Level of comparison*	*Data included in comparison*	*Data comparison*	*Choices included in comparison*
Comparison 1	Comparison of respondents	Data from all respondents to Questionnaires 1 and 2	Data from the 4 duplicate choices in Questionnaires 1 and 2	1, 4, 5 and 6
Comparison 2	Comparison of respondents	Data from respondents who completed Questionnaire 3 and were defined as reliable in Comparison 1	Data from all 12 duplicate choices in Questionnaires 1/2 and 3	1 to 12
Comparison 3	Comparison of respondents	Data from respondents who completed Questionnaire 3 and were defined as unreliable in Comparison 1	Data from 8 duplicate choices in Questionnaires 1 and 3 (excluding the 4 duplicate choices in Questionnaires 1 and 2)	2, 3, 7, 8, 9, 10, 11, 12
Comparison 4	Comparison of responses	Data from all respondents to Questionnaires 1/2 and 3	Data from all 12 duplicate choices in Questionnaires 1/2 and 3	1 to 12

Table 4 **Random effects probit models: summary of data included in each model**

Model	*Data included in model estimation*
Baseline	Data from questionnaires 1 and 2 but excluding data from immediate reliability checks (i.e. includes data from choices 1, 4, 5 and 6, and excludes data from choices 10, 12, 14 and 16)
Model 1	Data from questionnaires 1 and 2, including data from the immediate reliability checks (i.e. choices 10, 12, 14 and 16) but excluding original data from those checks (i.e. choices 1, 4, 5 and 6)
Model 2	Same data as in Baseline model but excluding data from respondents who did not respond to the postal survey
Model 3	Data from questionnaire 3 (postal questionnaire) only

Table 5 Identification of unreliable respondents: number (%) of respondents

	Comparison 1 (n=4 choices being compared)	*Comparison 2 (n=12 choices being compared)*	*Comparison 3 (n=8 choices being compared)*
Same response given to all duplicate choices	331 (56.6%)	46 (28.9%)	29 (31.5%)
Different response given to 1 duplicate choice only	179 (30.6%)	57 (35.8%)	31 (33.7%)
Different responses given to 2 duplicate choices only	58 (9.9%)	31 (19.5%)	22 (23.9%)
Different responses given to 3 duplicate choices only	10 (1.7%)	15 (9.4%)	9 (9.8%)
Different responses given to 4 duplicate choices only	7 (1.2%)	4 (2.5%)	0
Different responses given to 5 duplicate choices only	n/a	2 (1.3%)	0
Different responses given to 6 duplicate choices only	n/a	2 (1.3%)	0
Different responses given to 7 duplicate choices only	n/a	2 (1.3%)	1 (1.1%)
TOTAL	585 (100%)	159 (100%)	92 (100%)

Table 6 **Identification of unreliable responses: Comparison 4**

Questionnaire 1	**Questionnaire 2**		
	MRI	Arthroscopy	Total
MRI	1155	149	1304
Arthroscopy	168	829	997
Total	1323	978	2301

Table 7 **Identification of unreliable responses: Comparison 5**

Questionnaire 1/2	**Questionnaire 3**		
	MRI	Arthroscopy	Total
MRI	1965	177	2142
Arthroscopy	230	616	846
Total	2195	793	2988

Table 8 Random effects probit models: all data

Variable	Baseline			Model 1			Model 2			Model 3		
	Coeff.	95% CI: lower	95% CI: upper	Coeff.	95% CI: lower	95% CI: upper	Coeff.	95% CI: lower	95% CI: upper	Coeff.	95% CI: lower	95% CI: upper
COST	-0.00268	-0.00289	-0.00247	-0.00268	-0.00288	-0.00247	-0.00291	-0.00323	-0.00259	-0.00284	-0.00317	-0.00252
ARTH	0.147	-0.119	0.415	0.131	-0.134	0.397	0.259	-0.156	0.674	0.207	-0.219	0.635
TIME	-0.0289	-0.0333	-0.0244	-0.0276	-0.0320	-0.0232	-0.0329	-0.0397	-0.0260	-0.0320	-0.0388	-0.0251
RES	3.898	3.662	4.135	3.848	3.611	4.084	3.929	3.569	4.289	3.741	3.371	4.110
CONS	0.324	0.230	0.419	0.310	0.216	0.404	0.349	0.199	0.498	0.237	0.090	0.384

Model details:
Baseline: Data points = 6970; Respondents = 585; Observations/respondent (mean) = 11.91; $Chi^2 = 2081.64$ ($p<0.01$)
Model 1: Data points = 6938; Respondents = 585; Observations/respondent (mean) = 11.86; $Chi^2 = 2048.36$ ($p<0.01$)
Model 2: Data points = 2998; Respondents = 251; Observations/respondent (mean) = 11.94; $Chi^2 = 943.65$ ($p<0.01$)
Model 3: Data points = 3001; Respondents = 251; Observations/respondent (mean) = 11.96; $Chi^2 = 880.50$ ($p<0.01$)

Table 9 Comparison of individual respondent model coefficients: differences between coefficient values from models estimated using data from Questionnaires 1+2 and models estimated using data from Questionnaire 3

Variable	Mean difference	SD	95% CI for difference between means	Median difference	IQ Range
COST	0.009	0.13	-0.00796 to 0.0261	0	-0.0664 to 0.0664
ARTH	11.31	248.31	-19.5 to 42.2	0	-32.51 to 42.74
TIME	-0.23	5.29	-0.891 to 0.426	0	-1.10 to 0.77
RES	-4.55	114.12	-18.7 to 9.63	0	-36.90 to 36.71
CONS	-0.31	62.61	-8.10 to 7.47	0	-16.87 to 16.87

Table 10 Probit model to investigate factors associated with reliable respondents

Variable	Coefficient	Standard Error	p
Constant	-0.591	1.237	0.63
Age	0.047	0.030	0.11
Sex	0.146	0.196	0.45
Knee injury at initial data collection	-0.014	0.204	0.94
MRI at initial data collection	0.431	0.507	0.39
Arthroscopy at initial data collection	0.160	0.389	0.67
Believe in free health care	0.075	0.218	0.72
Active sports person	-0.021	0.238	0.92
Regional / International athlete	0.215	0.202	0.28
Recent knee injury at follow-up	-0.110	0.462	0.81
Recent MRI at follow-up	0.592	0.881	0.50
Recent arthroscopy at follow-up	-1.477	0.673	0.02
Unreliable respondent in Comparison 1	0.624	0.216	0.004

Log Likelihood: -120.30

FIGURE 1: COST ATTRIBUTE

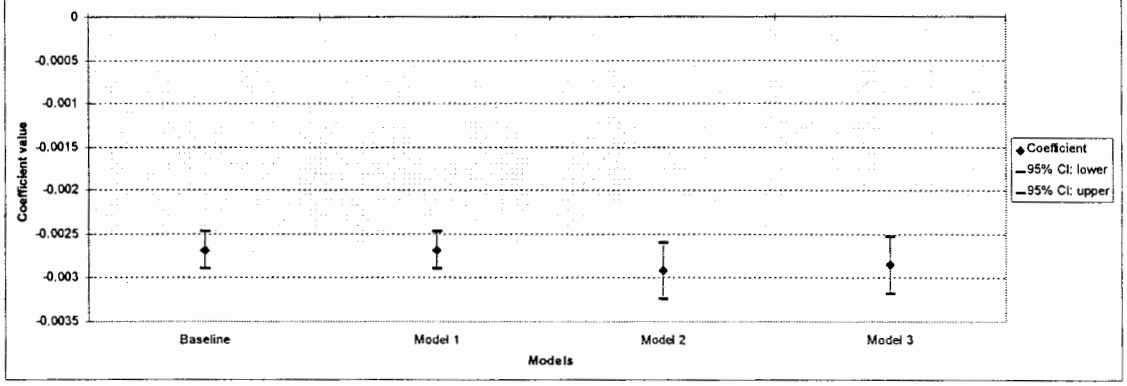

FIGURE 2: ARTHROSCOPY ATTRIBUTE

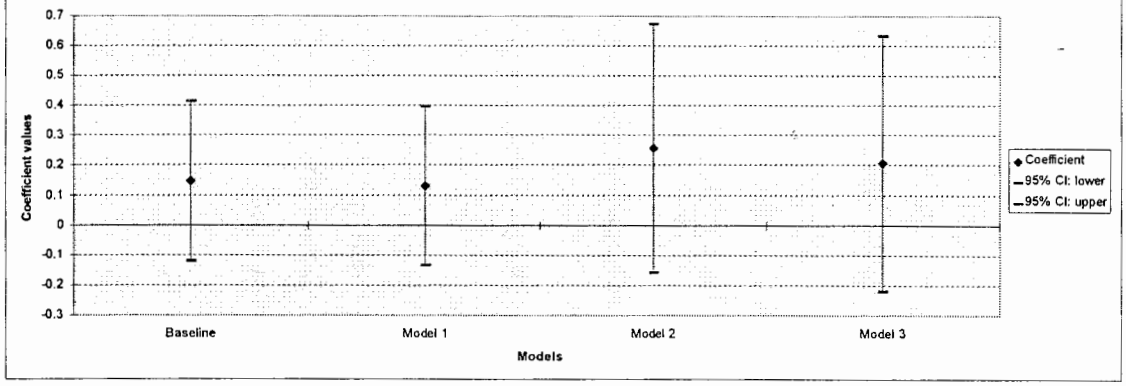

FIGURE 3: TIME ATTRIBUTE

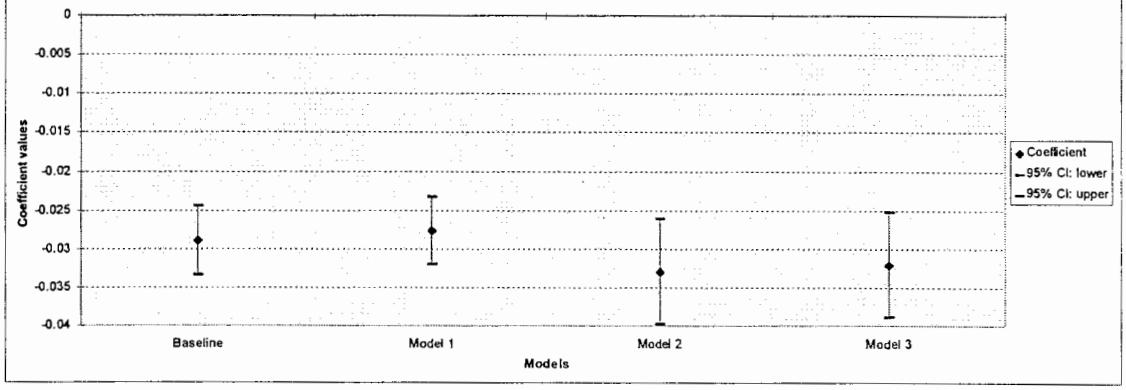

FIGURE 4: KNEE RESOLUTION ATTRIBUTE

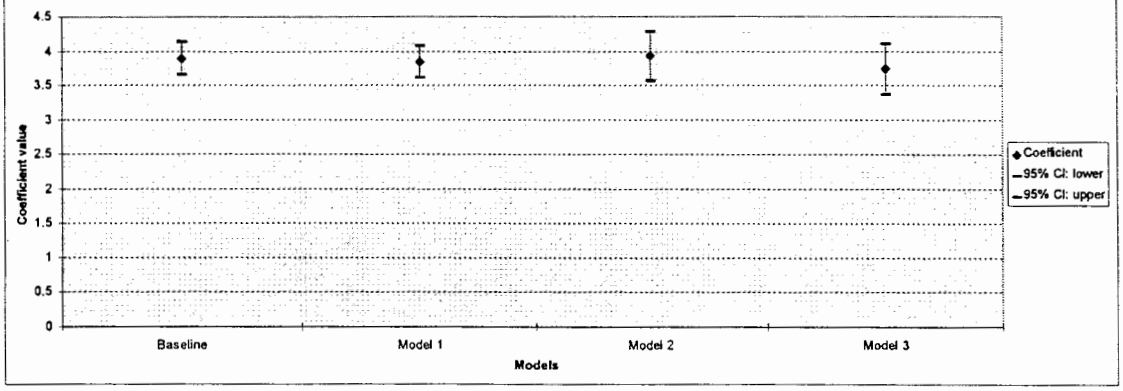

FIGURE 5: CONSTANT TERM