Environmental & Resource Economics (2006) 34: 45–50
DOI 10.1007/s10640-005-3783-x

Comment on "Revealing Differences in Willingness to Pay Due to the Dimensionality of Stated Choice Designs: An Initial Assessment"

F. REED JOHNSON
Research Triangle Institute, Research Triangle Park, P.O. Box 12194, 27709-2194, NC, USA (e-mail: frjohnson@rti.org)

Accepted 1 December 2004

Abstract. There are very few studies that quantify the interactions and tradeoffs between statistical and cognitive efficiency in designing stated-choice studies. While a conceptual framework for evaluating cognitive strategies would be desirable, Hensher adopts a strictly empirical approach in this experiment. The success of the study must be evaluated in light of his aggregating attributes rather than controlling the number of attributes, asymmetry in the narrow-range and wide-range attributes, and lack of orthogonality between the number of attributes and number of alternatives. Nevertheless, Hensher challenges uncritical acceptance of any given set of design features and correctly insists that we confirm our experience with rigorous, quantitative experiments.

Key words: choice experiments, experimental design, stated-choice methods, stated preference, transportation economics

JEL classification: C25, C93, R41

The conventional wisdom in stated-choice (SC) methods is that "everything matters." Elicitation format, amount and content of introductory information, placement of elements, face-to-face versus self-administration, as well as the list of design features included in David Hensher's design of designs (DoD) study, all seem to influence the way subjects evaluate preference-elicitation tasks (Arentze et al. 2003; Ben-Akiva et al. 1992; Blamey et al. 2000; Boyle et al. 2001; Bryan, et al. 2000; Darmon and Rouziès 1991; Louviere et al. 2003; Steenkamp and Wittink 1994). It is perhaps inevitable that this kind of research raises more questions than it answers. This paper is only one in a series of ambitious DoD studies Hensher is undertaking, so I am sure many of these questions eventually will be answered elsewhere.

1. Tradeoffs Between Statistical and Cognitive Efficiency

The extensive literature on statistically efficient design is not matched by a corresponding literature on cognitively efficient design. Moreover, there are very few studies that quantify the possible tradeoffs between statistical and cognitive efficiency (Arentze et al. 2003; Louviere 2001; Wand et al. 2001). It is quite possible that we could devise a D-optimal survey that ordinary subjects would find impossible to take. Understanding such tradeoffs would benefit from a conceptual framework for cognition. Such a framework would presumably yield testable hypotheses and help identify potential cognitive constraints that limit the quality and quantity of realistically obtainable preference information (Simonson and Tversky 1992; Slovic 1995). Ideally, understanding the cognitive strategies SC subjects use would also point the way to one or more indexes of cognitive burden that might help classify survey features and guide strategies to maximize overall survey efficiency.

In this paper Hensher takes a strictly empirical approach to this problem. He focuses on a number of survey features that all practitioners must define: number of attributes, number of levels, number of choice sets, and attribute ranges. Although all these features affect statistical efficiency, there is no analysis here of how the survey versions affect variance (Louviere et al. 2001, 2002). Apparently that is the subject of a separate paper. Rather, this paper tests whether various features separately or in combination affect willingness to pay (WTP) and value of travel time saved (VTTS). The results are mixed, and Hensher is prudent in refraining from drawing strong conclusions.

2. Number of Attributes and Aggregation

Constructing an experiment that varies the number of attributes without altering substitution relationships in the design is challenging. Hensher solves this problem by aggregating time and cost subcategories. By assuming the utility-weighted average preferences for the disaggregated attributes should equal the preferences for the aggregate attribute, Hensher claims to be able to isolate the effect of number of attributes. There are a couple of problems with this approach. First, as he admits, the experiment is really a test of attribute aggregation, rather than attribute number. Subjects who see the aggregate attribute may make different, unobserved assumptions about the composition of the aggregate, possibly based on their own experience. Regression results that indicate that predicted utility does not vary significantly with the degree of disaggregation do not really yield a generalizable conclusion about the possible effect of number of attributes.

In addition, the practical problem in designing SC surveys typically is not what level of aggregation to employ, but which and how many of a potentially long list of attributes to include. The experiment of interest, therefore,

would seem to be to test the consequence of including one or more relatively unimportant attributes in addition to a smaller set of relatively important attributes. Thus I fear that this feature of the study does not yield much insight for practitioners.

3. Range and Asymmetry

The range effect is the strongest result in the paper and is consistent with White et al. (1998). One important feature of the design is that narrow-range versions employ symmetric levels around the zero baseline, while wide-range versions employ asymmetric levels (the average level is greater than zero). The asymmetry was required because the design was pivoted-off of the features of an actual trip, leaving insufficient space to construct symmetric differences for the wide-range attribute. Hensher acknowledges the potential for differential framing effects apart from the difference in range. Regression results confirm the sensitivity of the results to this artifact, although the effect appears to be mitigated by controlling for the features of the reference trip. Again, it is not clear to what extent the seriousness of the range result is generalizable.

Constructing levels as percent differences relative to individual-specific reference-trip features has an additional problem. Each subject evaluates different absolute attribute levels, but preference estimates assume utility is invariant for the same percentage differences. This imposes implicit constant elasticity constraints on utility functions that seem unnecessary. To further complicate matters, subjects were shown absolute differences in the task screens.

4. Experimental Design

It is difficult to construct comparably efficient experimental designs with varying attributes, levels, and alternatives. Varying levels of statistical efficiency should, in principle, merely effect estimation precision. Unfortunately, attribute levels in practical experimental designs often are unbalanced. The resulting loss of orthogonality can induce collinearity and biased estimates (Kuhfeld 2004). Considering the relative simplicity of the setup, there are some notable imbalances in this design. For example, Table I compares

Table I. Combination counts of number of attributes and number of alternatives

	nalt = 2	nalt = 3	nalt = 4
nattr = 3	3	0	1
nattr = 4	1	2	1
nattr = 5	2	1	1
nattr = 6	0	2	2

counts for combinations of number of attributes and number of alternatives. The fewest number of alternatives never appears with the most number of attributes, while the fewest number of alternatives appears three times with the fewest number of attributes. These combinations leave room for only one design that combines the other two alternative levels with the smallest number of attributes. Such correlations might help explain why models that interact only one design feature yield some significant results, while models that control for all design features do not.

5. Preference Heterogeneity

Mean WTP is a primary goal of SC surveys, so it is natural to ask how design features might effect such estimates. However, mixed logit and hierarchical Bayes' methods now make it easy to explore how such features affect individual-level parameter estimates (Train and Sonnier 2004). Figure 1 shows an example from an experiment using two versions of a design with two different levels of attribute-level overlap or repetition within choice sets.[1] The figure shows the individual-level distributions of one preference parameter for the low and high overlap designs. While the means of these two distributions are different, the increased resolution in estimated taste heterogeneity from the high overlap design appears to be far more important.

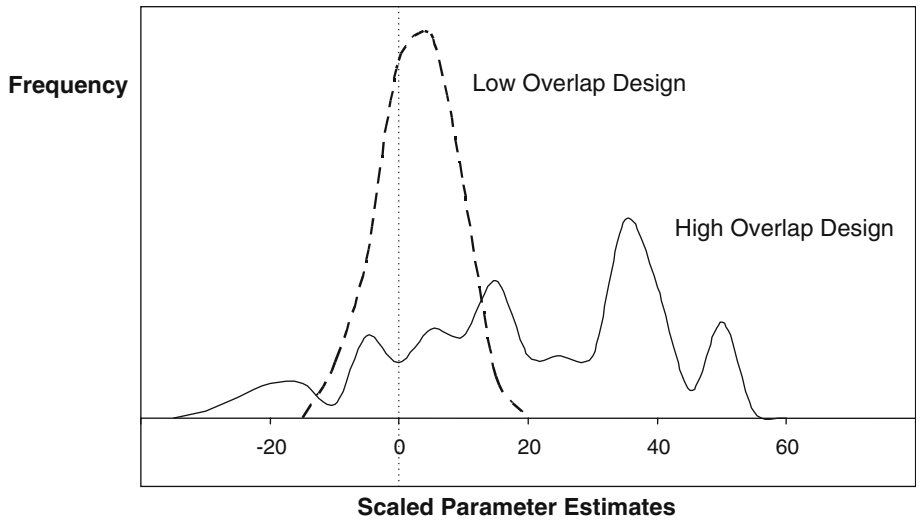

Figure 1. Distributions of a scaled individual-level parameter

6. Contribution of the Paper

Despite general acceptance of the idea that "everything matters," experienced SC researchers generally find a basic survey structure that seems to work well for a variety of applications. Hensher challenges uncritical acceptance of any given set of design features. He insists that we confirm our experience with rigorous, quantitative experiments. While we await additional findings from his research agenda, this paper leaves us with some uncertainty about the degree to which our preference estimates are induced, or at least influenced, by our favorite SC flavor. This paper sets a high standard for thinking creatively about how to quantify survey context effects. Nevertheless, other practitioners are likely to share my disappointment that Hensher does not provide more concrete counsel in this paper on how to improve the state of the practice. I look forward to seeing that counsel in future papers.

Note

1. This example is based on additional analysis of a study reported in Maddala et al. (2003).

References

Arentze, T., A. Borgers, R. DelMistro and H. Timmermans (2003), 'Transport Stated Choice Responses: Effects of Task Complexity, Presentation Format and Literacy', *Transport Research E* **39**, 229–244.

Ben-Akiva, M., T. Morikawa and F. Shiroishi (1992), 'Analysis of the Reliability of Preference Ranking Data', *Journal of Business Research* **24**, 149–164.

Blamey, R., J. Bennett, J. Louviere, M. Morrison and J. Rolfe (2000), 'A Test of Policy Labels in Environmental Choice Modeling Studies', *Ecological Economics* **32**, 269–286.

Boyle, K. J., T. P. Holmes, M. F. Teisl and B. Roe (2001), 'Comparison of Conjoint Analysis Response Formats', *American Journal of Agricultural Economics* **83**(2), 441–454.

Bryan, S., L. Gold, R. Sheldon and M. Buxton (2000), 'Preference Measurement Using Conjoint Methods: An Empirical Investigation of Reliability', *Health Economics* **9**, 385–395.

Darmon, R. Y. and D. Rouziès (1991), 'Internal Validity Assessment of Conjoint Estimated Attribute Importance Weights', *Journal of The Academy of Marketing Science* **19**, 315–322.

Kuhfeld, W. (2004), *Marketing Research Methods in SAS, TS-694*. Carey, North Carolina: SAS Institute.

Louviere, J. J. (2001), 'What If Consumer Experiments Impact Variances as Well as Means: Response Variability as a Behavioral Phenomenon', *Journal of Consumer Research* **28**, 506–511.

Louviere, J. J., D. J. Street and L. Burgess (2003), 'A 20+ Years Retrospective on Choice Experiments', in Wind Yoram and E. Green Paul, eds., *Marketing Research and Modeling: Progress and Prospects*. New York: Kluwer Academic Publishers.

Louviere, J., D. Street, R. Carson, A. Ainslie, J. DeShazo, T. Cameron, D. Hensher, R. Kohn and T. Marley (2002), 'Dissecting the Random Component of Utility', *Marketing Letters* **13**, 177–193.

Louviere, J., D. Hensher and J. Swait (2002), *Stated Choice Methods: Analysis and Application*, Cambridge, UK: Cambridge University Press.

Maddala, T., K. A. Phillips and F. R. Johnson (2003), 'An Experiment on Simplifying Conjoint Analysis Exercises for Measuring HIV Testing Preferences', *Health Economics* **12**(12), 1035–1047.

Simonson, I. and A. Tversky (1992), 'Choice in Context: Tradeoff Contrast and Extremeness Aversion', *Journal of Marketing Research* **29**(4), 281–296.

Slovic, P. (1995), 'The Construction of Preference', *American Psychologist* **50**, 364–371.

Steenkamp, J. -B. E. M. and D. R. Wittink (1994), 'The Metric Quality of Full-Profile Judgments and the Number-of-Attribute-Levels Effect in Conjoint Analysis', *International Journal of Research in Marketing* **11**, 275–286.

Train, K. and G. Sonnier (2004), 'Mixed Logit with Bounded Distributions of Partworths, Forthcoming', in A. Alberini and R. Scarpa eds., *Applications of Simulation Methods in Environmental and Resource Economics*. Kluwer Academic Publisher.

Wand, D., L. Jiuqun and H. Timmermans (2001), 'Reducing Respondent Burden, Information Processing Effort, and Incomprehensibility in Stated Preference Surveys: Principles and Properties of Paired Conjoint Analysis', *Transportation Research Record* **1768**, 71–78.

White, P. J., R. D. Johnson and J. J. Louviere (1998), *The Effect of Attribute Range and Variance on Weighted Estimates*. Department of Marketing, The University of Sydney (unpublished paper).