Environmental & Resource Economics (2006) 34: 173–188
DOI 10.1007/s10640-005-4817-0

What You Don't Know Might Hurt You: Some Unresolved Issues in the Design and Analysis of Discrete Choice Experiments

JORDAN J. LOUVIERE
Centre for the Study of Choice (CenSoC), Faculty of Business, University of Technology, Sydney, Australia (e-mail: deci@bigpond.net.au/jordan.louviere@uts.edu.au)

Accepted 2 November 2005

Abstract. The papers and comments in this issue focus on four broad areas related to understanding and modeling choices: (1) The use of laboratory experiments to improve valuation methods; (2) The design of stated preference choice set and choice occasions; (3) Latent class models as means of identifying and accommodating preference heterogeneity; and (4) Accommodating uncertainty about the "true" model, modeling ranking and rating tasks and pooling data sources. In what follows I offer some comments on each area, and briefly discuss several unresolved issues associated with each area, closing with some comments about future research opportunities.

Key words: discrete choice experiments, choice models, unobserved variability

1. The Use of Laboratory Experiments to Improve Valuation Methods

I will not only focus my discussion on the use of Discrete Choice Experiments (or DCEs, as they are now known in health economics), but will also comment on other types of experiments. DCEs involve designing choice options described by attribute level combinations. DCEs in environmental and resource economics typically involve two or more choices, but health economists often offer participants designed choice options one-at-a-time. More generally, however, DCEs involve constructing multiple scenarios/choice sets, from which participants choose one of the designed options or a non-designed option like a constant status quo, a status quo that varies across participants and/or an option to choose none of the options.

After working for more than a decade with environmental and resource economists, I still find it strange that many researchers and research problems are influenced by legal discussion and decisions. That is, it is not obvious that what is required to satisfy legal hurdles is good science, nor is it obvious that good science will satisfy legal hurdles. So, many issues raised by Harrison perplex me because I'm interested only in developing better approximations to how individuals make choices; I'm not interested in whether my research withstands legal scrutiny. For example, the external validity of DCEs likely rests on the degree to which a DCE simulates all key aspects of a real decision, including the incentive compatability of questions, framing of situations, contexts, consequences, etc. If one can design experiments that simulate real choice situations as closely as possible, one should be more likely to obtain results that mimic real life. If one fails to capture all the relevant aspects of a real situation, there should be systematic deviations. It's that simple.

It's also that complex. I'm not trying to trivialize the need to understand and do these things, but I find many experiments designed and implemented by economists to be unintuitive (I've participated in several to see for myself). So, few experimental economics studies simulate the types of markets that I study; most seem to simulate various auctions and/or dynamic markets like stock markets. The consumers I study do not "bid" for products or services – they are sold in traditional markets via retail or similar outlets (including online), and while "price negotiations" sometimes occur, they are uncommon in markets I study. For example, I rarely see sequential bids, except in housing auctions (common in Australia). Thus, it seems likely that many participants come to such experiments with little real auction experience, and despite efforts to "teach" them rules, etc., I am not convinced that naïve subjects placed in strange tasks and/or situations with often unintuitive (to me or the subjects!) payoffs and rules behave in ways that tell us much about real behavior. Instead, I think it may be time to ask if this tells us something about the psychology of experimental artifacts and/or the psychology of experimental economics artifacts.

Back to DCEs. A key advantage, and indeed the power of DCEs as I originally conceived them (e.g., Louviere and Woodworth 1983) is that they can simulate virtually all aspects of real markets as closely as desired, which means "subject to time and resources available". That is, given time, money and other resources, one can construct "real" stores or outlets that display all the key features of stores/outlets, including variations in product availability and prices. Indeed, "simulated stores" often are used to study and model uptake of new products or new product variants. When properly designed, such simulated shopping experiments accurately predict the actual market shares realized at and after launch/introduction.

The problem of modeling demand for radically new technologies, products and services has much in common with non-market traded goods because they do not exist, and typically will not exist for some time. So, one cannot expect samples of consumers to know about and/or have experience with the offerings. So, lack of knowledge and/or experience pose challenges for researchers. One solution involves "Information Acceleration Methods" or "Acceleration of Information Methods" (AIMs). Developed at MIT by Glen Urban and colleagues (e.g., Urban et al. 1990, 1996, 1997; see also Hoeffler 2003) in the late 1980s, AIMs rely on multimedia and other technologies to simulate the processes by which individuals become aware of new technologies/products, search for and acquire information about benefits and/or problem solutions, decide whether to consider them and whether they can take advantage of what they offer, decide if they want to buy a product now available, or wait to see how the product market develops and evolves over time.

Devinney, Louviere and Coltman (2004a, b, 2005) discuss designing and implementing AIMs quickly and easily by automating as much of the process as possible. They discuss case study applications involving voice-activated interactive PDAs, advanced commercial websites and new forms of space tourism. AIMs allow one to study how consumers become aware, search for and acquire information, and how this information impacts future choices. So, AIMs can simulate the market evolution stages of (1) becoming aware, (2) deciding what information to acquire and how to acquire it, (3) deciding whether and which options to consider, (4) deciding whether to choose now, delay or never choose, and (5) if choosing now, deciding which option(s) to choose.

So, several of Harrison's issues relate to needing to understand how real markets evolve and change, and how consumers in those markets learn and make decisions as both consumers and markets evolve. So, some of these issues can be resolved by careful attention to how real markets for the goods of interest could or should establish themselves and evolve over time. For example, we recently studied the demand for space tourism offerings here in Australia (Zero-G, Suborbital and Orbital trips). We developed multimedia materials to inform and educate subjects, and collected data from high net worth individuals over the web. Many of the attributes related to how these offerings could change over time and when options would become available.

Once one understands how markets and consumers are likely to evolve, AIMs can be used to simulate the process. This implies that researchers must invest as much effort as possible in advance of fieldwork to design and implement realistic simulations. Ultimately, one should expect that more realistic market simulations will win out over less realistic experiments, because it's hard to see how one can learn much about consumer behavior in real markets without studying these markets. To summarize, unless one is relatively certain that subjects understand the goods, understand the market for them, understand the context, etc., it is unlikely that a DCE will be incentive-compatible, will provide accurate and unbiased estimates of the tradeoffs and choices, etc.

Having said that, I agree with Harrison's discussion of the potential usefulness of combining experimental and real market data. I'll have more to say about this later, including potential reasons why both revealed and stated preference models do not transfer and/or cannot be cross-validated.

2. The Design of Stated Preference Choice Sets and Choice Occasions

I applaud the idea of using designs that differ in complexity to estimate the same effects, and in so doing obtain information about the impacts of various design components on choice variability. I also like the idea of pivoting designs around respondents' current (most recent) choice because this adds realism. Unfortunately, adding realism is not a statistical design property; hence, it is unclear how statistically efficient such pivot procedures are, particularly if "fiddling" is done to minimize dominance and/or eliminate "unrealistic" options. Statistical design theory now exists to optimize the efficiency of designs for generic conditional logit models, and alternative–specific conditional logit models (see e.g., Burgess and Street 2003, 2005a, b; Street and Burgess 2004; Street et al. 2001, 2005, forthcoming). Thus, one can determine how statistically efficient approaches like Hensher's (this volume) are. Statistical efficiency matters because a design that is 40% efficient "throws away" 60% of the observations. Now that one can calculate the efficiency of any conditional logit design or collection of conditional logit designs, such calculations should be required for publication. Specifically, the level of efficiency of one's design(s) should be provided, and the actual design itself should be provided for others to check them. For example, in Hensher's case, this would entail checking all the choice sets comprising each experiment (i.e., all persons' sets – the experiment was individualized), as well as the collectivity of choice sets for all experiments to check the obtained statistical efficiency.

The idea of considering the effect of properties of choice designs on parameters estimated from the designs is potentially interesting and important. However, it is critical that a master design is itself able to estimate the effects of interest. If one is interested only in attribute main effects, it *may* be sufficient to use a main effects plan (i.e., a design of resolution 3 in the statistical literature). If one also must estimate two-factor interactions, one needs a design of resolution 5. Res 5 designs can be quite large; e.g., if there are five attributes each with four levels, one needs at least $4^4 = 256$ designs specified by a master design. If only some main and interaction components (effects) are of interest, degrees of freedom considerations imply that fewer designs need to be specified by the master design. Unfortunately, however, we do not know how to construct these smaller designs in general.

So, one must check the properties of designs used in a master design, which cannot be achieved without access to the exact designs, and in the case of computer generated designs like Hensher's, one needs all the data. It also is worth noting that computer-generation does not insure that the resulting designs will have the properties claimed. Thus, one needs to verify claimed properties, and to do that one needs access to designs and datasets. CenSoC research teams have reviewed many designs in marketing and applied economics, and we conclude that researchers generally seem unaware that it is important to disclose all relevant details of designs so they can be checked against theoretical benchmarks. Design disclosure should be a publication requirement, and researchers should recognize that the statistical properties of designs greatly matter. Moreover, all designs are not equal, and as I later show, results from fractional designs are likely to be biased. So, one's results depend heavily on choice of design(s), which implies that design properties and proper design selection underlie and are confounded with results from different designs. Researchers should recognize that the designs chosen for DCEs are at least as, if not more important than, the models that one uses to analyze the resulting data.

Like Hensher, CenSoC research teams also are studying the effects of design efficiency on "respondent efficiency" (a term Valerie Severin coined in her PhD thesis, U. of Sydney, 1999). For example, we recently designed 22 experimental conditions based on different paired comparison designs (full design details available from me). Each design differed in numbers of attributes (4, 8, 12 or 16) and numbers of pairs (8, 12, 16, 32). We studied two product categories (delivered pizzas, island holidays); attributes in each category were based on Severin's results, such that the first four attributes had the largest effects, and so on. We designed experiments that systematically varied in statistical efficiency, numbers of attributes and numbers of attribute differences in each pair.

We used covariance heterogeneity models (CHM) to estimate variance scale ratios associated with each condition relative to a reference condition (e.g., DeShazo and Fermo 2002; Swait and Adamowicz 2001a, b). Results are described in more detail in a CenSoC Working Paper; we found that error variances increased as logarithmic functions of numbers of attributes and design efficiency (variance = inverse of scale) in both categories. Whereas previous researchers focused on the number of attribute differences, this measure is a function of numbers of attributes and design efficiency, not an antecedent.

3. Using Latent Class Models to Identify and Capture Preference Heterogeneity

Economists and marketers seem to be overly obsessed with preference heterogeneity, which is only one potential source of unobserved variability, and may not always be a large source. That is, there are many other sources of unobserved variability; and arguably one should pay at least as much as if not more attention to them instead of focusing on heterogeneity to the exclusion of virtually all else (see, e.g., Louviere 2001a, 2004a, 2004b; Louviere et al. 2005 forthcoming). I now briefly discuss several major issues in these latter references.

1. Stochastic Utility Components (i.e., "errors") not only are not IID, they are systematically related to what is manipulated in DCEs and/or characteristics of choosers. "Errors" are not unidimensional, suggesting that components of variance models should better approximate the process than most popular choice models. Indeed, it is surprising that only Cardell (1997) seems to have considered variance components for choice models. So, researchers should acknowledge that there can be many unobserved effects, few of which are associated with preference heterogeneity. They also should acknowledge that error variances systematically vary with several factors (see, e.g., Louviere 2001a, b, 2004a, b), which also implies that there should be large differences between models estimated in utility space and willingness-to-pay (WTP) space (Train and Weeks 2005, forthcoming). WTP space is a transform of the utility space that involves expressing all estimates as ratios of price coefficients (the price estimate is $-1/$(error standard deviation)). If errors are IID, there should be no difference in the fits of the two models except for rounding errors; otherwise, there should be large differences. That is, the WTP specification cancels scale, which makes empirical sense only if errors are IID. Train and Weeks tested this notion, and found large differences. They also noted that the distributions assumed for random effects in Mixed Logit (McFadden and Train 2000) are not inconsequential, and ratios of some distributions are non-sensical. Indeed, far too many DCE papers treat analysis as merely a statistical exercise, devoid of behavioral theory. So, the field would benefit from placing more emphasis on conceptual and behavioral frameworks to motivate and guide model specification instead of merely fitting complex statistical models.

That brings me to latent class models (LCMs). LCMs also confound various sources of unobserved variability, but at least are relatively easy to estimate, and can potentially avoid meaningless WTP ratios. They also seem to fit data at least as well as random parameter models, so until we better understand how to deal with potential scale confounds and develop better behavioral theory, LCMs deserve more attention, especially in applied economics.

I'm optimistic that we will find ways to overcome scale issues and develop better insights into sources of unobserved variability. For example, CenSoC teams recently implemented 66 experimental conditions involving 1200 web-based survey respondents to study the effects of numbers of attributes, levels and choice options on the ability to estimate choice models for single subjects. If one can estimate models for single individuals, random coefficient or finite mixture models would not be needed because (by definition) one can directly estimate the empirical distribution of preferences. This ability also would allow tests for other sources of unobserved variability as one can hold variability in preferences constant. Frankly, I look forward to getting past the notion that preference heterogeneity = unobserved variability, and addressing more interesting questions about what other sources of unobserved variability matter and how to best account for them.

2. The combinatorics of DCE responses. For example, suppose one asks three people to respond to four scenarios describing public buses that vary in fare, service frequency and travel time for a certain origin/destination, as shown below. Each person says whether they would/would not use each bus described. Such a DCE has 2^4 possible response patterns (i.e., yes/no combinations) because one can say yes/no (2) to each description (4). Below are two possible 1/2 fractions of the 2^3 factorial that can be combined to yield the full factorial of three factors, each varied over 2 levels.

Scenario (combination)	Fare	Service frequency	Travel time	Decision Rule 1	Decision Rule 2	Decision Rule 3	Total Yes's
Fraction 1							
1	Low	Low	Low	No	No	Yes	1
2	Low	High	High	No	No	Yes	1
3	High	Low	High	No	No	No	0
4	High	High	Low	No	Yes	No	1
Fraction 2							
1	Low	Low	High	No	No	Yes	1
2	Low	High	Low	Yes	Yes	Yes	3
3	High	Low	Low	No	No	No	0
4	High	High	High	No	No	No	0

Now, consider three decision rules for "saying yes": (1) Say yes if fare is low, travel time is low and service frequency is high; (2) Say yes if service frequency is high and travel time is low; and (3) Say yes if fare is low. These rules are shown in shown in columns 5, 6 and 7 in the table above; the last column contains the number of yes's resulting from the three rules. Each fraction has 16 yes/no combinations, all eight scenarios (the factorial) represent 256 combinations, but only about half of them make sense due to preference directionality.

Decision Rule 1 leads to a "yes" only in Fraction 1, illustrating that what one "sees" and analyzes in DCEs depends on designs used. Decision Rules 2 and 3 produce two observations of "yes". I analyzed each rule by fitting a linear probability model (LPM) to the responses, using effects codes and their cross-products. LPMs simply summarize the marginal frequency (conditional means) of yes's. The results are in the table below.

Linear Probability Model Results

Effect	Fraction 1 R1	Fraction 1 R2	Fraction 1 R3	Fraction 2 R1	Fraction 2 R2	Fraction 2 R3	Both Fractions R1	Both Fractions R2	Both Fractions R3
Constant		0.250	0.500	0.250	0.250	0.500	0.125	0.250	0.500
Fare		0.250	−0.500	−0.250	−0.250	−0.500	−0.125	0.000	−0.500
Service		0.250	0.000	0.250	0.250	0.000	0.125	0.250	0.000
Time		−0.250	0.000	−0.250	−0.250	0.000	−0.125	−0.250	0.000
F×S							−0.125	0.000	0.000
F×T							0.125	0.000	0.000
S×T							−0.125	−0.250	0.000
F×S×T							0.125	0.000	0.000

For Fraction 1, Rule 1 produces all no's; Rule 2 produces an incorrect sign for fare when the effect should = 0; Rule 3 produces the "right" sign for fare. For Fraction 2, Rules 1 and 2 produce the SAME statistical results (Rule 2 again has an incorrect fare effect), and Rule 3 again gives the "right" fare sign. If we combine both fractions, we can "see" that the rules really differ, and indeed we can "see" that fare = 0 for Rule 2. This example shows that it is dangerous to make inferences about how humans make decisions using fractional designs.

More generally, it is *NOT POSSIBLE* to learn about decision rules unless one uses complete factorials or all possible choice sets. That is, each response pattern observed in a fractional design is observationally equivalent to many different decision rules. Thus, researchers need to understand that most statistical choice models are only paramorphic representations based on linear utility functions (Hoffman 1960), and in many, if not most, cases it is highly likely that the estimated models are seriously incorrect (biased).

To show that such models are incorrect, I fit binary logit models to the choices implied by the three decision rules, as shown below. Note that (a) decision rules are deterministic, (b) each rule is different, and (c) "error" only results from ignorance of true rules. All models have the "correct signs" for all estimates, a "sniff test" many researchers use to decide if models make intuitive sense. Obviously, this is inadequate. Notice that model fits vary a lot depending on the fraction; fraction two produces much higher fits. Because we know the decision rules, we also know the data generating process is deterministic and non-additive for at least two rules (trivially additive for the third). This example should discourage researchers who claim that they can identify segments/classes with different decision rules. Without full factorials or all possible choice sets, this is a meaningless exercise, and such claims should not be made.

Binary Logit Model (Data Aggregated Over All 3 Rules)				
Effect	Fraction 1	Fraction 2	Both Fractions (main effects only)	Both Fractions (all effects)
Constant	−2.821	−3.224	−1.566	−3.147
Fare	−2.127	−8.979	−1.566	−5.428
Service	2.127	3.224	1.001	2.801
Time	−2.127	−3.224	−1.001	−2.801
F×S				−0.173
F×T				0.173
S×T				−2.801
F×S×T				0.173
Pseudo ρ^2	0.231	0.854	0.511	0.620

Indeed, a simple thought experiment shows that claims of capturing different decision rules in almost all DCEs are unfounded. Consider a typical DCE involving 16 choice sets (scenarios) with an associated yes/no response. The number of possible response patterns is $>65{,}000$, so very few will be observed in any realistic sample. Now let the DCE have four choice options in each of 16 choice sets; the possible response patterns $=4^{16}$, or around 4.3 billion, putting to rest any claims of capturing decision rules in DCEs.

Equally important, these "simple" examples suggest that editors and reviewers should ask researchers who claim to "capture" distributions of utility estimates to explain to the research community how one can say anything meaningful with typical sample sizes used in academic or commercial research. Indeed, the examples also suggest that one needs to be *very lucky* to obtain reliable estimates of a distribution of preferences. If this is not enough to discourage researchers from trying to do this, here's more bad news: virtually all the choice patterns are *inconsistent* with linear, additive utility functions (e.g., for 8 scenarios, only 6/256 are). So, the likelihood of obtaining unbiased estimates from models that assume linear, additive utility specifications without interactions is very small, which can be seen in the LPM results (both fractions yield estimates inconsistent with those associated with the full factorial).

3. Generalizing from ONE dataset and ONE model. The above discussion implies formidable challenges in specifying suitable models of individual decision/choice processes; but things actually are worse than that. Specifically, any one model represents a sample of *size one* of possible models that could be considered; and that model typically is estimated from *ONE* data set, which is a sample of *size one* of possible datasets that can be analyzed. This implies that one needs theory and/or reliable empirical evidence to guide research if one wants to generalize. "Better" theory will not come from statistics or statistical models; it will come from careful consideration of fundamental, substantive behavior(s) and creative insights into individual and group behavioral processes.

Meyer et al. (1997) noted this generalization problem, and conceptualized it as follows:

Y = f(X, Z, C, G, T), where
Y = a vector of behavioral outcomes;
X = a matrix of factors describing options/outcomes;
Z = a matrix of individual/group factors;
C = a matrix of contex/environmental factors;
G = a matrix of geographical/spatial factors; and
T = a matrix of time-varying factors.

One's ability to generalize depends on how well "f" is specified in the above expression and/or the extent to which the components of each vector are constant (or nearly so) across empirical domains. It also implies that we must include as many possible components of variance in the stochastic utility component as possible (e.g., Cardell 1997). That is, the more constant components in any given dataset, the harder it will be to generalize models estimated from that *ONE* dataset. It will also be harder to generalize complex models that include latent terms as one needs to generalize "extra" latent parameters and/or account for differences in datasets to the extent that they manifest themselves as parameter and/or variance differences. Thus, editors and reviewers should ask those who claim to estimate segments, parameter distributions, etc., to provide evidence of ability to generalize as a publication requirement. Whenever possible, DCE models should be tested against real market behavior(s) and/or validity tests should be designed to maximize the chance that models will fail in certain ways to allow one to discriminate models. Many statistical choice models now in vogue are likely to over-fit data and capitalize on chance; moreover, as Dawes and Corrigan (1974) noted more than 30 years ago, the types of utility functions typically estimated insure high fits even when seriously wrong. So, research would benefit from better behavioral theory that can be used to guide specification of models to be estimated from DCE choice data instead of merely fitting this or that statistical model to data because others have done so.

4. Accommodating Uncertainty About the "True" Model, Modeling Ranking and Rating Tasks and Pooling Data Sources

The preceding section notes that few decision rules that consumers use in DCEs are likely to be strictly additive. Ranking tasks can provide extra preference information to improve statistical efficiency, but they cannot yield insights into "true models" or individual decision rules without full factorials. Moreover, subjects can use different cognitive processes to rank sets of options. For example, Finn and Louviere (1992) proposed a choice task and associated model called Best–Worst Scaling (BWS) or "Maximum Difference Scaling". BWS is a theory about how individuals choose, respectively, best and worst options from sets; BWS assumes that individuals choose that pair of options in each set that differ the most on an underlying latent dimension. Marley and Louviere (2005, forthcoming) show how different ways of making such choices imply different models, and each model has different measurement properties, and in some cases, different degrees of bias.

CenSoC teams are using best–worst questions to get more preference information from DCEs. That is, subjects report most + least preferred options and/or report additional most and least preferred options from the remaining options, until some/all options are ranked. Louviere et al. (2004) show how to combine sequences of best–worst questions (BWQs) with efficient DCE designs to estimate a model for each person. BWQs can rank options in each choice set, but BW rankings may differ from directly ranking *n* options from 1 to *n*. Also, the choice variability associated with ranking from 1 to *n* may not be the same as with repeated BW choices. Previous research suggests bias in 1 to *n* rankings even after scale differences in ranking depths are taken into account (e.g., Ben-Akiva et al. 1992; Hausman and Ruud 1987). The earlier decision rule examples suggest why one should expect bias ACROSS individuals, namely different people use different rules, and most rules not only are not additive, they may not be independent of errors.

One does not need a complete ranking to obtain extra preference information. Consider a choice set of four options, from which a person chooses, respectively, their most and least preferred options. If the set consists of {a,b,c,d}, and the ordering of preference is {1,2,3,4}, then the person should say that a is most preferred and d is least preferred. This implies that we also know a > b, a > c, a > d, a > bc, a > bd, a > cd, a > bcd, b > d, and c > d. There are 11 non-empty, non-singleton subsets, namely all subsets containing at least two options. Information about nine comes from only one set of BW choices. Humans tend to respond more consistently to extreme options, with response inconsistency increasing towards mid-ranked options (e.g., Louviere 2001a, b).

That brings me to ratings. I'm not sure why economists are interested in ratings, other than it may be that ratings give more information. The measurement properties of ratings are controversial, and many believe that ratings are biased, at best providing order, not metric information. Indeed, it is unclear how to interpret rating tasks unless one assumes an ordered preference measure conditional on making a choice (e.g., Louviere 2001b). I often show my classes that one typically gets very different WTP estimates from "would/would not" choose responses compared with rating responses (in DCEs). Typically, rating task estimates of WTP are too large to be credible, particularly compared to choice-based estimates of WTP.

On the other hand, I agree with Layton and Lee (this volume) that researchers should try to obtain more preference information and pool data to obtain more efficient estimates. I noted earlier, however, that one also should try to free up variance components for identification that are constant in some data sources but not in others; and it's unclear what ranking and rating responses have to contribute to that issue. I'll close this discussion by saying that it is not sufficient to "assume" that certain response tasks and associated responses produce data that have thus and such properties. Instead, one needs a theory of the data generation process used by the individuals to produce the data, and many possible data generation processes can underlie an individual's ranking or rating of options. The field would benefit from postulating a theory of the data generation process, and developing rigorous ways to test if predictions from the theory and/or the assumptions that underlie the theory are satisfied. We now have such a theory for best and worst judgments, and we know that such judgments are cognitively fairly easy for individuals. We also know the properties of resulting measures if the theory satisfies the data from the judgments (choices). So, I urge researchers to pay more attention to processes that can underlie DCE responses, and not simply view the problem as "the response data are of thus and such form, and so I need to analyze them with such and thus statistical models that have been previously been proposed and applied, or this new model that I can write down."

5. Where Is/Should the Field Be Going?

I suggested several potentially fruitful future research areas, including developing behavioral theory and using it as a basis for formulating and testing models, recognizing and developing ways to capture random error components, developing ways to test how well model results can be generalized, understanding and modeling how DCEs impact the behavior of subjects, providing evidence that model assumptions are satisfied, etc. I noted that some researchers are trying develop ways to estimate models for single individuals and to develop and estimate models that can capture systematic relationships between random component variances and covariances and the attributes of choice options and characteristics of individuals. So, more recognition that scale plays a fundamental behavioral role in choices would be welcome.

I would like to see more cross-disciplinary collaboration among researchers interested in understanding and modeling choices. Despite some progress, it remains true that most research is what Kuhn (1965) described as "normal science in a paradigm", yet virtually all major scientific breakthroughs result from cross-disciplinary innovations. For example, the level of knowledge and understanding of statistical design theory exhibited in most published DCE papers is very low and/or exhibits significant errors. Similarly, the level of knowledge and understanding of key concepts and issues in cognitive psychology or judgment and decision making outside of psychology is generally low. Not surprisingly, the same can be said of psychologists with respect to design theory or microeconomic concepts. So, a little knowledge can be a dangerous thing, and I urge researchers to seek out suitable, well-trained colleagues in cognate fields to be collaborators instead of trying to "learn it yourself". Doing the latter well is a hard ask, and typically is associated with significant gaps in understanding, although I readily admit that there can be exceptions.

I doubt that more complex statistical choice models are the answer to the empirical issues that the field faces. Instead the way forward lies in development of better behavioral theory and insights to guide model development and empirical research. I look forward to that future.

6. Concluding Comments

DCEs and associated choice models have come a long way in a short period of academic time, but this research is in its infancy, with many unresolved problems and issues. I discussed some of these unresolved issues and problems, and tried to show that some issues are not worth further research time because they are inherently unresolvable, like the idea that one can "really" understand decision processes except in small, restrictive circumstances.

It may seem that my comments are largely negative, but I did not intend them to be. I'm actually optimistic that we will find ways to resolve many of these issues, and I anticipate great progress. I'm particularly encouraged by the fast-growing interest in DCEs by applied economists, who inevitably will make significant advances. Indeed, only 15 years ago there were few applied economists and only a small number of marketers and others interested in these problems. Now, it is fair to say that there are many more applied economists interested in DCEs than researchers in marketing and transportation.

The fact that a special issue like this appears is a testament to the notion that this research area has well and truly arrived in economics. There are many more academic economists than academic marketers and transport researchers, so I expect economists to dominate this area sometime in the coming decade. In the future this special issue may come to be seen as the tipping point from which economists went on to research leadership in this field.

References

Ben-Akiva, M., T. Morikawa and F. Shiroishi (1992), 'Analysis of the Reliability of Preference Ranking Data', *Journal of Business Research* **24**(2), 149–164.

Burgess, L. and D. Street (2003), 'Optimal Designs for 2^k Choice Experiments', *Communication in Statistics – Theory and Methods* **32**,2185–2206.

Burgess, L. and D. Street (2005a), 'Optimal Designs for 2^k Choice Experiments', *Communications in Statistics –Theory and Methods* **32**,2185–2206.

Burgess, L. and D. Street (2005), 'Optimal Designs for Choice Experiments with Asymmetric Attributes', *Journal of Statistical Planning and Inference*, **134**, 288–301.

Cardell, N. S. (1997), 'Variance Components Structures for the Extreme Value and Logistic Distributions with Applications to Models of Heterogeneity', *Econometric Theory* **13**(2), 185–213.

Dawes, R. M. and B. Corrigan (1974), 'Linear Models in Decision Making', *Psychological Bulletin* **81**,95–106.

David F. Layton and S. Todd Lee (2006). "Embracing Model Uncertainty: Strategies for Response Pooling and Model Averaging." This volume.

DeShazo, J. R. and G. Fermo (2002), 'Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency', *Journal of Environmental Economics and Management* **44**,123–143.

Devinney, T. M., J. J. Louviere and T. Coltman (2005), 'Utilizing Rich Multimedia Methods for the Elicitation of Preferences for Radical Future Technologies', /EMAC/ANZMAC/ / Joint Conference/, Milan, May.

Devinney, T. M., J. J. Louviere and T. Coltman (2004a), 'Utilizing Rich Multimedia Methods for the Elicitation of Preferences for Radical Future Technologies', /ESOMAR Conference in Marketing: Where Science Meets Practice, Oct, Warsaw, http://www.esomar.org/main.php?a = 3&p = 1161, pp. 271–288.

Devinney, T. M., J. J. Louviere and T. Coltman (2004b), 'Decision States and Information Acceleration', ANZMAC Conference, Wellington, Dec.

Finn, A. and J. J. Louviere (1992), 'Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety', *Journal of Public Policy and Marketing* **11**(1), 12–25.

Hausman, J. A. and P. A. Ruud (1987), 'Specifying and Testing Econometric Models for Rank-Ordered Data', *Journal of Econometrics* **34**,83–104.

Hoeffler, S. (2003), 'Measuring Preferences for Really New Products', *Journal of Marketing Research* **40**,406–420.

Hoffman, P. J. (1960), 'The Paramorphic Representation of Clinical Judgement', *Psychological Bulletin* **57**,116–131.

Louviere, J. J. (2001a), 'What if Consumer Experiments Impact Variances as Well as Means: Response Variability as a Behavioral Phenomenon', *Journal of Consumer Research* **28**(3), 506–511.

Louviere, J. J., R. T. Carson, A. Ainslie, T. A. Cameron, J. R. DeShazo, D. Hensher, R. Kohn, T. Marley and D. Street (2002), 'Dissecting the Random Component of Utility', *Marketing Letters* (Special Issue on the UC Berkeley Invitational Choice Symposium) **13**(3), 177–193.

Louviere, J. J. (2004a), 'Random Utility Theory-Based Stated Preference Elicitation Methods: Applications In Health Economics With Special Reference To Combining Sources of Preference Data', *CenSoC Working Paper No. 04-001*, http://www.business.uts.edu.au/censoc/papers/index.html.

Louviere, J. J. (2004b), 'Complex Statistical Choice Models: Are the Assumptions True, and If Not, What Are the Consequences?', *CenSoC Working Paper No. 04-002*, http://www.business.uts.edu.au/censoc/papers/index.html.

Louviere, J. J., L. Burgess, D. Street and A. A. J. Marley (2004), 'Modeling the Choice of Single Individuals by Combining Efficient Choice Exeriment Designs with Extra Preference Information', *CenSoC Working Paper No. 04-005*, http://www.business.uts.edu.au/censoc/papers/index.html.

Louviere, J. J., R. J. Meyer, D. S. Bunch, R. Carson, B. Dellaert, W. M. Hanemann, D. Hensher and J. Irwin (1999), 'Combining Sources of Preference Data for Modelling Complex Decision Processes', *Marketing Letters* **10**(3), 187–204.

Louviere, J. and G. Woodworth (1983), 'Design and Analysis of Simulated Consumer Choice or Allocation Experiments: an Approach based on Aggregate Data', *Journal of Marketing Research* **20**,350–367.

Marley, A. A. J. and J. J. Louviere (2005), 'Some Probabilistic Models of Best, Worst, and Best-Worst Choices', *Journal of Mathematical Psychology*, forthcoming.

McFadden, Daniel, Kenneth Train (2000) 'Mixed MNL Models for Discrete Response,' *Journal of Applied Econometrics*, **15**, 447–470.

Street, D., D. Bunch and B. Moore (2001), 'Optimal Designs for 2^k Paired Comparison Experiments', *Communications in Statistics – Theory and Methods* **30**,2149–2171.

Street, D. and L. Burgess (2004), 'Optimal and Near-Optimal Pairs for the Estimation of Effects in 2-Level Choice Experiments', *Journal of Statistical Planning and Inference* **118**,185–199.

Street D., L. Burgess and J. J. Louviere (2005), 'Quick and Easy Choice Sets: Constructing Optimal and Nearly Optimal Stated Choice Experiments', *International Journal of Research in Marketing*, forthcoming.

Swait, J. and W. Adamowicz (2001a), 'Choice Complexity and Decision Strategy Selection', *Journal of Consumer Research* **28**,135–148.

Swait, J. and W. Adamowicz (2001b), 'Choice Environment, Market Complexity and Consumer Behavior: A Theoretical and Empirical Approach for Incorporating Decision Complexity into Models of Consumer Choice', *Organizational Behaviour and Human Decision Processes* **86**(2), 141–167.

Train, Kenneth and Melvyn Weeks (2005), 'Discrete Choice Models in Preference Space and Willingness-to-Pay-Space'. in A. Alberini and R. Scarpa, eds., *Applications of Simulation Methods in Environmental Resource Economics* (pp. 1–17). Dordrecht, The Netherlands: Springer, Chapter 1.

Urban, G. L., J. R. Hauser, W. J. Qualls, B. D. Weinberg, J. D. Bohlman and R. A. Chicos (1997), 'Information Acceleration: Validation and Lessons from the Field', *Journal of Marketing Research* **34**,143–153.

Urban, G. L., J. R. Hauser, J. R. Roberts and H. John (1990), 'Prelaunch Forecasting of New Automobiles', *Management Science* **36**(4), 401–421.
Urban, G. L., B. D. Weinberg and J. R. Hauser (1996), 'Premarket Forecasting of Really New Products', *Journal of Marketing* **60**(1), 47–60.