Journal of Choice Modelling, 3(3), pp 57-72
www.jocm.org.uk

Discrete Choice Experiments Are Not Conjoint Analysis

Jordan J Louviere[1,*] Terry N Flynn[1,†] Richard T Carson[2,Ŧ]

[1] Centre for the Study of Choice, University of Technology, Sydney, PO Box 123 Broadway, Sydney, NSW 2007, Australia
[2] Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508

Received 25 April 2010, revised version received 31 October 2010, accepted 31 October 2010

Abstract

We briefly review and discuss traditional conjoint analysis (CA) and discrete choice experiments (DCEs), widely used stated preference elicitation methods in several disciplines. We pay particular attention to the origins and basis of CA, and show that it is generally inconsistent with economic demand theory, and is subject to several logical inconsistencies that make it unsuitable for use in applied economics, particularly welfare and policy assessment. We contrast this with DCEs that have a long-standing, well-tested theoretical basis in random utility theory, and we show why and how DCEs are more general and consistent with economic demand theory. Perhaps the major message, though, is that many studies that claim to be doing conjoint analysis are really doing DCE.

Keywords: discrete choice experiments, conjoint analysis, random utility theory

* Corresponding author, T: +61-2-95149799, F: + 61-2-95149897, jordan.louviere@uts.edu.au
+ T: +61-2-95149804, F: + 61-2-95149897, terry.flynn@uts.edu.au
Ŧ T: +1-858-534-3383, F: +1-858-534-7040, rcarson@ucsd.edu

ISSN 1755-5345

1 Introduction

Interest in the use of stated preference (SP) theory and methods has increased dramatically in agricultural and food economics, environmental and resource economics and health economics since the mid-1990's. SP methods are used to elicit an individual's preferences for "alternatives" (whether goods, services, or courses of action) expressed in a survey context. Thus, this class of methods for preference elicitation differs from traditional economic approaches which are based on revealed preference (RP) data obtained by observing individual behavior in real markets. This paper defines, compares and discusses two paradigms that are being used more and more widely in applied economics, and shows why one of them (conjoint analysis) generally is inappropriate for economic evaluation and should be used with caution in economic applications.

There are many ways to elicit stated preferences from individuals. We concentrate on two general paradigms for preference elicitation that have evolved over the past thirty plus years and that have substantial empirical records that can be assessed. These paradigms are "conjoint analysis" (CA) and "discrete choice experiments" (DCEs). Here we use the term DCE rather than the more commonly used term choice experiment to avoid confusion.[1] Academics and practitioners often seem to confuse both paradigms based on our review of the literature and considerable experience with them over the past 20+ years. Indeed, we believe that many researchers who claim to apply conjoint analysis really are using DCEs. In turn, this suggests that many researchers currently working on SP issues do not realize that both paradigms differ, much less how they differ. In fact, the two approaches differ substantially, and several of the ways in which they differ have significant implications for economic evaluation and related applications. Thus, the purpose of this paper is to discuss differences in both paradigms and explain why one should be cautious about using CA in many economics applications. We achieve this purpose by reviewing both paradigms, which involves describing and comparing both, noting differences and similarities. We then discuss how and why differences and similarities matter to researchers interested in applying these methods.

2 What is Conjoint Analysis?

Conjoint Analysis is a generic term used to describe several ways to elicit preferences. CA's origins are in psychology, principally associated with research dealing with ways to mathematically represent the behavior of rankings observed as an outcome of systematic, factorial manipulation (i.e., known as "factorial designs") of independent factors (also known as "attributes"). CA methods rely on formal proofs about the mathematical (algebraic) representations of rank orderings of orthogonal arrays (originally, complete factorial arrays). In particular, it can be shown (e.g., Krantz and Tversky 1971) that if a person ranks a full factorial design, their implied preferences can be represented "as if" they integrate and combine the factor levels in certain algebraic ways, such as "adding" or "multiplying" marginal preferences for each factor level to rank all factorial treatments (a treatment is a combination of factor

[1] Carson and Louviere (forthcoming) note that the term choice experiment has different meanings in other disciplines such as biology and physics and that nothing restricts choices to be "discrete" as commonly assumed in SP work.

levels). Specifically, a person's ranking of a set of factorial treatments must satisfy certain conditional and joint preferential independence and dependence properties in order to represent the ranking as one of the foregoing simple algebraic processes. Thus, CA evolved out of the theory of "Conjoint Measurement" (CM), which is purely mathematical and concerned with the behavior of number systems, not the behavior of humans or human preferences. That is, the axioms underlying the theory are purely mathematical, and posed challenges in early attempts to operationalize and apply the theory. Specifically, the rankings of real individuals generally violate one or more axioms, posing issues as to how many and which violations are too many and/or are sufficient to accept one model over another.

Two main features of CM are important for our discussion. The first is that the axioms of CM have some relationship to utility theory. Unfortunately, however, this relationship is very restrictive, such that CM has been superseded by more general standard neoclassical utility theory (e.g., Varian 1992) and its variants like prospect theory (e.g., Kahneman and Tversky 1979) that allow particular types of deviations from CM theory. Second, there is no error theory associated with CM, statistical or otherwise, which allow the theory to be represented as testable statistical models; hence learning that preference data do not correspond to simple conjoint models contributed to the development of statistical error theories in psychology.[2] Yet, these statistical error theories are not used in CA analysis nor do they underlie DCEs.

The CA literature is most extensive in academic and applied marketing, where it is widely used to solve practical marketing research problems (e.g., Green and Rao 1971; Cattin and Wittink 1982; Wittink and Cattin 1989). Much of the use of the term "conjoint" in various other applied microeconomics fields seem to stem from beliefs that they emulate what marketers do when collecting preference data with surveys. As CA applications began to expand rapidly in marketing in the late 1960's and early 1970s, one began to "see" many practical "enhancements" of the formal "Conjoint Measurement" theory (e.g., Krantz and Tversky 1971; Michell 1990), such as the following:

- Using fractional factorial arrays instead of complete factorial designs to elicit preferences for combinations of factor (attribute) levels;
- Using rating scales instead of rankings to elicit preference orders and/or utility differences;
- Using statistical models to analyze elicited preference responses and estimate preference parameters (i.e., part-worth utilities); and
- Using "choice simulators" to predict individual choices from sets of options and aggregate choices over samples of people.

The above list is not exhaustive; it is intended merely to be representative of some enhancements arising out of practical experiences. For example, it omits some analytical developments, such as multidimensional scaling, two-limit Tobit models, ordered logit or probit models and other censored and limited dependent variable regression techniques that match particular censoring properties of various types of CA preference responses obtained from ranking or rating of attribute combinations (e.g., Anderson and Bettencourt 1993; Harrison et al. 2005; Jedidi and Zhang 2002; Vriens et al. 1998). Indeed, the latter enhancements were largely statistical because

[2] One of the best known statistical error theories is Functional Measurement, which is the empirical implementation of Information Integration Theory (e.g., Anderson 1962, 1970).

they arose from considering ways to analyze certain types of response outcomes as data types. That is, few analytical developments in CA were derived from theoretical considerations about decision making processes; and most developments in CA (i.e., post-1975) have been ad hoc statistical and methodological "enhancements", virtually none of which were derived from underlying behavioral theory. For example, there are so-called "partial profile" methods (e.g., Bradlow et al. 2004), so-called "hybrid conjoint" methods (e.g., Green and Srinivasan 1978) or more recently, hierarchical Bayes and latent class methods used to estimate, respectively, continuous or finite mixture distributions of preferences (e.g., DeSarbo et al. 1992; Moore et al. 1998; Park 2004).

More generally, all CA methods have the following in common:

- CA research studies typically begin with what usually are ad hoc and researcher-specific ways to identify attributes. Identification methods/approaches range from various types of direct questions asking people what attributes drive their preferences (e.g., Louviere 1988) to relatively sophisticated, quasi-theoretical methods like Kelly's Repertory Grid (e.g., Timmermans et al. 1982). Thus, there is no accepted "standard" way to identify attributes; and to our knowledge, there is no consensus about how to identify attributes, other than possibly a consensus that one should identify attributes. It is worth noting that this stylized fact also applies to DCEs. It is worth noting that this step frequently is done by using qualitative research involving in-depth interviews and/or focus groups with the target population.
- Once attributes are identified, a researcher must assign them levels/values to express a range of actual or potential variation in the context being studied. Like attribute identification, there is little consensus about how to "do" this; hence, practice varies widely. After identifying attributes and levels, attribute levels are combined to describe alternative goods/offers. Attribute level combinations typically are generated by some type of experimental design, most often an orthogonal fractional array, although this need not be the case, as exemplified by work in Social Judgment Theory/Policy Capturing (e.g., Adelman et al. 1975). More generally, almost all published applications used low resolution designs with few attribute level combinations, the vast majority of which are so-called orthogonal main effects plans (OMEPs). OMEPs impose serious restrictions on one's ability to understand decisions and valuation processes as they require researchers to assume that preference or utility functions are strictly additive (i.e., full preferential independence). Perhaps more seriously, one cannot test the validity of this assumption and use more appropriate model specifications if the assumption is wrong because higher order interaction effects are deliberately confounded with lower order main effects to reduce the number of attribute level combinations, and hence, the number of questions. This matters because estimates of consumer surplus or WTP can exhibit large differences between incorrect additive forms and correct non-additive specifications.
- After designing an experiment, the next step is to design and implement a preference elicitation task based on the design. Again, there is little consensus about how to "do" this, resulting in wide variation in practice. For example, the literature reveals serious disagreements about how many attributes, attribute levels, attribute level combinations, etc., can and/or should be used in any particular task. There is much overlap in the traditional CA analysis and DCE

literatures, with most of the work done in DCE applications (e.g., Carson et al. 1994; Louviere et al. 2008).

- After designing a task, it has to be implemented, which involves sample selection and data collection. Again, there is little consensus on this, except perhaps that it is difficult to implement CA via telephone interviews. We return to sample selection issues later, but for now we note that many traditional CA studies potentially suffer from sample selection bias which may influence the interpretation of results.
- After collecting data, the analyst must analyze them. Again, practice differs widely with researcher preferences, training and predispositions. Moreover, such practices seem to be changing with the advent of computerized design, administration and analysis methods for traditional CA. We observed little consensus about proper way(s) to analyze and/or model CA preference data for particular applications.

The above leads to the conclusion that CA is what CA researchers do. CM theory originally was not a theory about the behavior of preferences or choices, but instead a theory about the behavior of sets of numbers in response to factorial manipulations of factor levels. However, if one can elicit numbers representing preferences from a person, and the numbers satisfy certain axiomatic and/or statistical conditions, CM theory tells us that one can represent the numbers "as if" the individual used a certain algebraic process to combine preferences for each level of each attribute into a preference for holistic combinations of attribute levels. For example, suppose one describes a health insurance policy by attributes like deductable amount, co-payment premium, which items/conditions are covered and to what extent, etc. CM theory can be applied if individuals have potentially different degrees of preference for each level of each of these attributes and if they combine the preferences for each attribute level into an overall preference for each insurance policy option.

Thus, traditional CM theory provides a way to study how such preferences are formed and identify processes individuals use to form them, but one rarely ever sees CA methods used to examine and model preference process(es) per se. Moreover, mainstream CA evolved away from full ranking of profiles towards these directions (e.g., Green et al. 2001): (a) assigning ratings to individual profiles one at a time, (b) self-explicated rating of attributes and their importance and (c) various types of adaptive models involving ratings of pairs of alternatives, often on the basis of a partial set of the attributes.[3] What is interesting about all of these "innovations" is that all depart from collecting data that were potentially consistent with neoclassical utility theory (ranking a full set of complete profiles). That is, they collect data in ways that cannot be analyzed to be consistent with neoclassical economic theory because ratings and attribute importance measures do not readily translate into choice or matching (e.g., direct expression of WTP) data, the primitives on which utility theory is based.

There is a tendency in the academic and practitioner literatures in marketing, as well as in some other literatures, to call any and all preference elicitation procedures involving some variation of attributes and levels "conjoint analysis". However, this makes the term "conjoint analysis" virtually meaningless because it does not denote or connote anything other than to imply any preference elicitation procedure that involves variation of attributes and levels. During the 1980s it became common to

[3] The exception is what Green et al. (2001) refer to as choice based-conjoint analysis (DCEs in this paper). We discuss DCEs at length below.

suggest that DCEs were "just another form of conjoint analysis" by calling them "choice-based conjoint analyses". However, this is potentially very misleading. The RUT basis of DCEs is very different from CM. Indeed, recognition of this distinction led health economists conducting DCEs to stop calling them conjoint analyses in the late 1990s; both main health economics journals (*Health Economics* and *Journal of Health Economics*) have now acknowledged this naming convention. The factorial experiment was borrowed from statistics and is standard fare; so it cannot be the core of what CA means. Most ways used to elicit preference information in contemporary CA are not unique to CA, and those that are unique tend to generate data that do not readily conform to standard neoclassical economic theory. Thus, we argue that in spite of common usage, it is a mistake to think of DCEs as a special case of CA.

3 What is a DCE?

In contrast to traditional CA that relies on CM, which is not a behavioral theory (of choice), DCEs are based on a long-standing, well-tested theory of choice behavior that can take inter-linked behaviors into account. The theory was proposed by Thurstone (1927), and is called random utility theory (RUT). Recent work in DCE theory and methods relies heavily on work by McFadden, who extended Thurstone's original theory of paired comparisons (pairs of choice alternatives) to multiple comparisons (e.g., McFadden 1986; McFadden and Train 2000; McFadden 1974; Thurstone 1927).[4] Unlike CM, random utility theory provides an explanation of the choice behavior of humans, not numbers.

Specifically, RUT proposes that there is a latent construct called "utility" existing in a person's head that cannot be observed by researchers. That is, a person has a "utility" for each choice alternative, but these utilities cannot be "seen" by researchers, which is why they are termed "latent". RUT assumes that the latent utilities can be summarized by two components, a systematic (explainable) component and a random (unexplainable) component. Systematic components comprise attributes explaining differences in choice alternatives and covariates explaining differences in individuals' choices. Random components comprise all unidentified factors that impact choices. Psychologists further assume that individuals are imperfect measurement devices; so, random components also can include factors reflecting variability and differences in choices associated with individuals and not choice options per se. More formally, the basic axiom of RUT is:

$$U_{in} = V_{in} + \varepsilon_{in}, \quad (1)$$

where U_{in} is the latent, unobservable utility that individual *n* associates with choice alternative *i*, V_{in} is the systematic, explainable component of utility that individual *n* associates with alternative *i* and ε_{in} is the random component associated with individual *n* and option *i*.

[4] It is useful to note that McFadden's (1986) *Marketing Science* article, which focused on data collected by the then rapidly expanding number of surveys used to collect CA data, raised many of the issues that we do. In particular, McFadden noted the highly restrictive nature of the CM framework, the advantage of the RUT framework, the need to take account of the error structure, the importance of possible divergences between SP and RP contexts, ways to incorporate various types of self-explicated preference information on attributes into the latent component of RUT models instead of using it to directly indicate preferences, and difficulties in using the extra information in rating data in a theoretically consistent manner.

Because there is a random component, utilities (or "preferences") are inherently stochastic as viewed by researchers. So, researchers can predict the probability that individual *n* will choose alternative *i*, but not the exact alternative that individual *n* will choose. RUT leads to families of probabilistic discrete choice models that describe how choice probabilities respond to changes in choice options (or equivalently, their attributes) and/or covariates representing differences in individual choosers. Thus, the probability that individual *n* chooses option *i* from a set of competing options is:

$$P(i|C_n) = P[(V_{in} + \varepsilon_{in}) > Max(V_{jn} + \varepsilon_{jn})], \text{ for all j options in choice set } C_n. \quad (2)$$

All terms were defined earlier, except for Max, the maximum operator. Equation 2 says that the probability that individual *n* chooses option *i* from the choice set C_n equals the probability that the systematic and random components of option *i* for individual *n* are larger than the systematic and random components of all other options competing with option *i*.

Different probabilistic discrete choice models can be derived from equation 2 by making different assumptions about probability distributions for ε_{in}, such as assuming the random components are distributed as non-independent, non-identically distributed normal random variates (Thurstone also considered restricted cases like IID normal). In contrast, McFadden assumed the random components were IID Gumbel (Extreme Value Type 1). Unfortunately, Thurstone's non-IID normal distribution assumption retarded development of RUT and delayed development of practical multiple choice models because the normal distribution lacks a closed form for more than two choice options. Indeed, until recently, good approximations for solving multiple integrals to compute choice probabilities were unavailable (e.g., McFadden and Train 2000; Yellott 1977). The Gumbel distribution closely resembles the normal but is slightly asymmetric; it has the advantage of yielding closed form expressions for the choice probabilities if random components are IID, namely the well-known multinomial logit (MNL) model (also known as a conditional logit model) still used in practical applications. If random components are not IID, Gumbel distributions also do not give closed-form expressions for the choice probabilities. The non-IID case has spawned relatively new ways to estimate choice models, such as simulated maximum likelihood or hierarchical Bayes. Thus, RUT accommodates different distributional assumptions that lead to different probabilistic discrete choice models. Specific distributions have particular properties, and currently much work is focused on statistical methods for distinguishing between competing specifications.

4 Why RUT is a Comprehensive Behavioral Theory

We now introduce an abstraction of the overall process by which individuals come to be aware that options exist and evolve into subsequent actions/states leading to a researcher observing their market behavior(s) (Figure 1). We can rely on RUT to give a comprehensive conceptualization of the entire system and/or any of its substates.[5] Relationships implied by Figure 1 comprise a series of conditional probability events that can be approximated in a variety of ways, such as a nested, sequential choice process that ultimately leads to a set of discrete choice options. It also should be

[5] Of course, RUT has many well-known limitations. See Hensher et al. (2005) for a useful discussion.

obvious that individuals can choose more than one option at a time or over time. For example, a person can choose to buy several types of breakfast cereal at the same time or over different purchase occasions. The person can choose to consume (use) each in different quantities (volumes) and can choose to purchase them more or less often over time (different inter-choice or usage periods). The latter behaviors can be viewed as volumes or usage rates of options; in turn, they are conditional on prior choices. One can view Figure 1 as a series of inter-related choices that consumers can (and typically do) make; these inter-related choices can be modeled within a RUT framework.

CA traditionally has been used to model only one level in the hierarchy in Figure 1, and has not been used to model inter-related levels in a comprehensive hierarchy of choices. Moreover, without a comprehensive, overarching behavioral theory, it is unclear how one could use CA to develop sound theoretical models of the processes implied by Figure 1. In fact, it is worth noting that CA adherents in marketing and transport largely focus on the final-stage process of choosing choice alternatives denoted by the dashed line enclosing part of Figure 1 that is labeled "Traditional CA Territory". Thus, traditional CA ignores prior processes and states either because there is little awareness that they exist or matter except descriptively (e.g., Howard and Sheth 1969) and/or there is no behavioral theory underlying CA to deal with them. The emphasis, if not the bias, in CA related research has been predicting outcomes instead of understanding decision processes. This emphasis is unfortunate because it is well known that one can predict well (at least in the short term like the "next quarter") without basic theory or understanding of process (e.g., Anderson and Shanteau 1977; Dawes and Corrigan 1974; Louviere 1988). Many academics and practitioners appear to equate prediction success with understanding an underlying behavioral process, although they are clearly different; of course, each has its place.

Many DCE applications resemble traditional CA simply because they use survey questions about combinations of attribute levels. The combinations of attribute levels typically are generated by an experimental design, and samples of people respond to them one-at-a-time. Respondents typically are asked to rate the alternatives on a scale or to rank them. Of course, when two alternatives are rank ordered this can be viewed as a binary choice; and it is well-known that rank order data can be expanded into several sets of implied discrete choices (e.g., Chapman and Staelin 1982). Further, one can avoid assuming cardinality of ratings data by simply using the greater than, equal to, or less than information implicit in ratings. Nonetheless, neither CA nor DCEs should be seen simply an elicitation format or way to combine SP questions with associated experimental designs. The key difference in traditional CA and DCEs lies in the critical role of error components. For CA, treatment of error components is an afterthought, whereas in DCEs, it is the starting point.

Another difference in traditional CA and DCEs is a link to the economic concept of demand based on utility maximization under some type of constraint (e.g., budget or time). Here the link with DCEs is direct and easy to establish given its RUT foundation and its ability to link different stages of the decision making process. A DCE choice set always includes at least one feasible alternative.[6] In contrast, traditional CA sometimes offers respondents an entire set of infeasible alternatives. For instance, it would be an acceptable CA task to ask people to rank order automobile

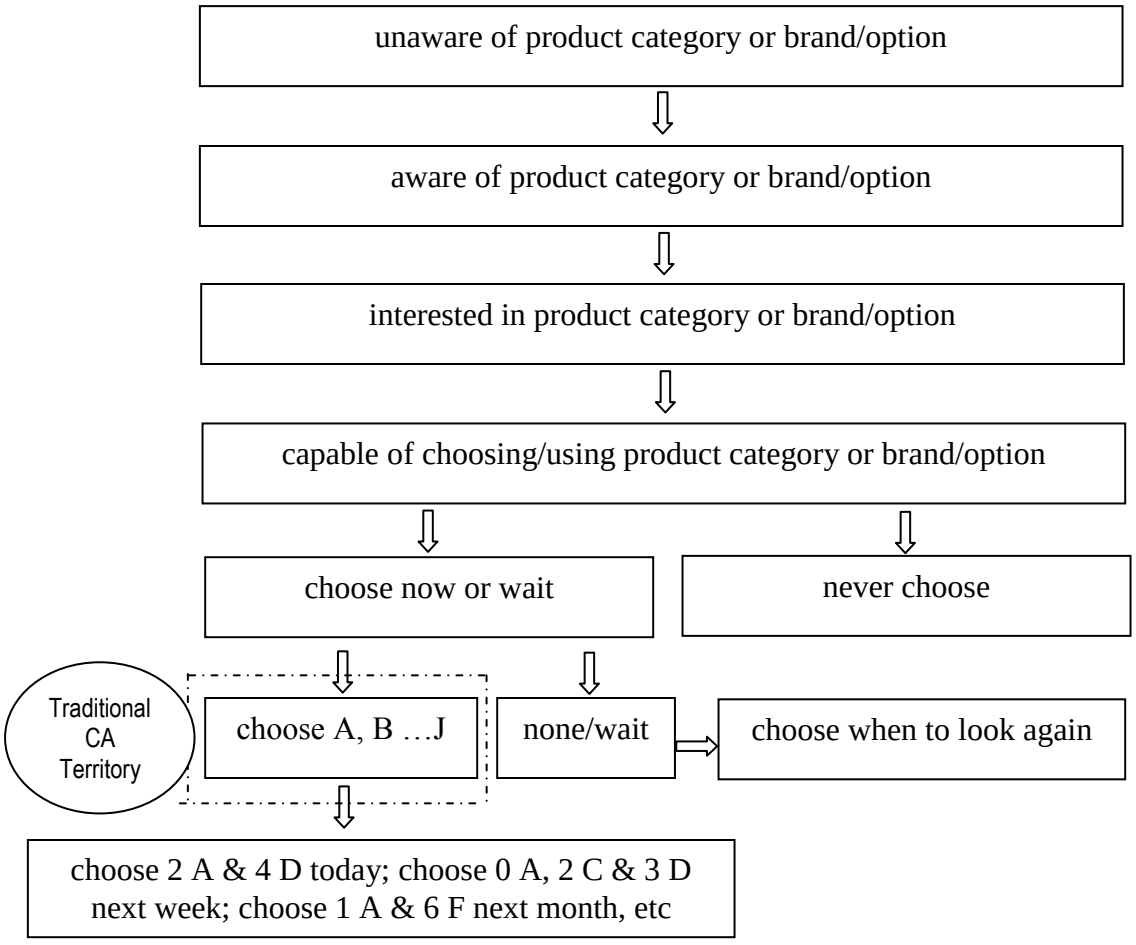
[6] This often is a status quo or no purchase alternative; but, of course, some contexts may not have a status quo option equivalent to no purchase. For example, parents may be legally required to choose a particular school for their child, and a grocery store may offer the choice of paper, plastic, or a person's own bag for groceries.

preferences for a Ferrari, a Hummer, and a Rolls Royce even if they could not afford to purchase any of them, which would be a meaningless DCE question. The opposite variant of this issue may be more important in practice. That is, CA (and some DCE) studies often screen out individuals who have not previously purchased a product that represents one alternative in the study. Although this (largely) avoids infeasibility problems, one then must deal with the issue that some increases in product sales from improving one or more attributes (including potentially lowering the cost) may come from people who do not currently buy the product. This poses no conceptual difficulty for the RUT framework as shown in Figure 1 below.

5 On the External Validity of CA and DCEs

The underlying behavioral theoretical foundation of DCEs is RUT, which allows one to deduce the relationship between discrete choice models estimated from DCEs and corresponding choices that would likely be made in a different context, such as offers made in an actual market. Perhaps the best way to see the difference between traditional CA and RUT is to note that a change in the magnitude of the variance of the error term under CA has no influence on preferences and choices because CA is

Figure 1: Decision-making process of a respondent



not concerned about error variances per se. This is not true under RUT, and an extreme case occurs when the error component variance becomes arbitrarily large relative to the systematic component. When this happens, the market shares of all alternatives become approximately equal. Likewise, as the error component variance becomes small relative to the systematic component, the chance of some alternatives being chosen becomes arbitrarily close to one or zero.

Equation 1 allows for the distribution properties of random components to be as general as desired. However, there is an important identification restriction that must be imposed (e.g., Ben-Akiva and Morikawa 1990). Any particular vector of estimated choice model parameters ("betas") is inversely proportional to the size of the variances of associated random components. The restriction does not matter in predicting choice probabilities, but one must set the "scale" of the utility vector to some constant for identification purposes when making inference about the latent scale. Indeed the focus on odds ratios in epidemiology and health services research may be responsible for misunderstanding of the importance of the variance scale confound in RUT-based tasks (McCabe et al. 2006) recently noted by Flynn et al. (2008).

Model parameters and random component variances are inversely proportional for any probability distribution assumed for random components. So, if one defines a constant, say λ, to be a scalar multiple of the true vector of utility parameters, then in general $\lambda = k(1/\sigma_\varepsilon)$, where k is a constant of proportionality, and σ_ε is the standard deviation of the random component (error distribution). Due to the inverse relationship between error variance and scale, we say that λ "scales" the estimated vector of beta parameters; that is

$$V_i = \lambda\beta_k X_{ki} + \varepsilon_i. \qquad (3)$$

V_i is the systematic utility component, λ is a scale parameter, β_k is a K (=1, 2,…, K) element vector of utility estimates (parameters); X_{ki} (i=1, 2,…, I) is a K x I element array of attribute effects (a so-called "design matrix") and ε_i is the random component associated with the *i*-th choice option..

The inverse relationship between scale and random component variance holds for all choice models; so, all parameters estimated from choice data consistent with RUT have this inherent identification issue.[7] In DCEs, this scale confound suggests great caution in comparing model parameters for individuals, groups, conditions, contexts, etc. That is, small random component variances give rise to larger estimated model parameters and larger variances give rise to smaller estimates. Consequently, even if a preference generation process in two groups or contexts is the same, if the random component variances differ, the magnitudes of model parameter estimates will seem to differ. Thus, all comparisons of RUT-based choice models must take differences in random component variances into account or risk incorrect and misleading statistical conclusions (e.g., Fiebig et al. 2010; Swait and Louviere 1993).

The parameter estimate-scale confound has profound implications for modeling many choices, and starkly differentiates DCEs from traditional CA. For example, Ben-Akiva and Morikawa (1990) reasoned that if the underlying preference process was the same for stated preference (SP) and revealed preference (RP) data sources, the two vectors of estimated model parameters should be proportional. That is, rewrite

[7] Differences in random component variances for different individuals (or groups) also manifest themselves in other ways in parameter and/or standard error estimates in traditional CA models, which does not seem widely recognized.

equation (3) for SP and an RP data sources, with *I* and *J* total choice options, respectively, as follows:

$$\text{SP:} \quad V_{iSP} = \lambda_{SP}\beta_k X_{kiSP} + \varepsilon_{iSP}, \tag{4a}$$

$$\text{RP:} \quad V_{jRP} = \lambda_{RP}\beta_k X_{kjRP} + \varepsilon_{jRP}. \tag{4b}$$

If the two vectors of βs are the same in each data source but the scales (λs) differ, the SP betas will be proportional to the RP betas, and the constant of proportionality will be $\lambda_{SP}/\lambda_{RP}$. Ben-Akiva and Morikawa tested this expectation in two data sets related to preferences for interurban travel in the Netherlands and could not reject the hypothesis that the βs were proportional. Since then, researchers in different disciplines have tested this proposition dozens of times, with the general result being that it often holds to a close first approximation (e.g., Louviere and Eagle 2006; Louviere et al. 2000; Louviere et al. 1999).

Without an underlying behavioral theory one is unable to deduce the relationship between traditional CA parameters or model predictions and real market or RP behaviors. Thus, only RUT-based DCE preference elicitation methods have testable (and well-tested) theoretical links with real behavior(s). CA lacks a sound, theoretical relationship with real market choice behavior(s), which serves to reinforce the ad hoc, predominantly statistical and methodological nature of CA research and practice. It also explains why most so-called empirical "validity tests" in the CA literature merely represent cross-validation tests with hold-out samples. Naturally, such "tests" have nothing whatsoever to do with real "external" validity, and merely measure test-retest reliability and prediction shrinkage. Real external validity tests must compare trade-offs made in SP elicitation tasks with trade-offs made in revealed choice data from real markets or in different SP environments. Of course, even worse are so-called "hit rate tests", which are nothing more than percent correct predictions of first choices, a criterion with no known statistical properties.

Error components in SP and RP environments often differ, and it is not surprising that "noise" levels in RP environments often can be quite large. That is, extraneous factors often compete for consumer attention, and it may well be in the interests of some firms to make straightforward comparisons more difficult. Indeed, in many cases random component variances should be smaller for SP than RP data because SP data are obtained under fairly well-controlled circumstances where individuals focus on fairly specific decision and choice tasks that encourage them to ignore omitted factors and/or assume that they are constant. When this is the case, one should expect (see equation 4) preference parameters estimated from SP data to be larger (in absolute magnitudes) than similar parameters estimated from RP data.[8] These scale differences have implications for predicted market shares even if the deep preference parameters are equivalent in RP and SP data. This issue lies at the heart of why it is difficult to use and validate estimates from traditional CA studies. If one wants to make accurate predictions using SP data in an RP context where there are multiple choice options, one typically must rescale the SP estimates, yet CA has no mechanism to achieve this because it has no real error theory.

Another problem for traditional CA methods is that they do not naturally yield willingness to pay (WTP), the Hicksian consumer surplus measure that economists and many other applied researchers like to work with. There are two reasons for this.

[8] Clearly, this generalization need not always hold. For instance, if respondents in a particular DCE exercise devote little attention to the questions, SP parameter estimates can be noisier.

First, traditional CA does not start from the position of a well-defined status quo option. As already mentioned, an acceptable CA task would be to ask people to rank order preferences for a Ferrari, a Hummer, and a Rolls Royce even if they cannot afford to purchase them. Second, the scale factor has an influence on statistics related to WTP. For some statistics, like marginal WTP for a change in an attribute level, the scale factor drops out in some simple specifications where the ratio of the two parameters is the appropriate measure. This happy circumstance often is used in an ad hoc manner in CA studies even though it does not naturally follow from the statistical specification fit to the data. More generally, WTP measures depend on scale estimates, which is particularly true for studies that attempt to estimate total WTP for a good rather than a marginal change in a good. That is, a respondent's certainty in their decision (as one, but not the only, characterization of variance scale) almost certainly will change when asked about larger changes than smaller ones. This raises the key question of to what market context does one "normalize" the WTP estimate? One school of thought is to normalize to the RP context, but the opposite argument also can be made. That is, particularly in a public policy context, the less noisy SP context is a better reflection of what well-informed consumers would do.

6 On the Versatility of RUT-Based SP Methods

Traditional CA methods depend on orthogonal or near orthogonal arrays of attribute level combinations as ways to sample profiles from full factorial arrays of attribute levels, but RUT-based SP methods do not have this limitation. There are many ways to elicit stated preferences that one can use to estimate discrete choice models that are consistent with RUT but inconsistent with CA. A few simple examples should suffice to make this point. First, consider a task in which respondents are shown a set of possible choices and then asked which ones, if any, they would prefer relative to the current status quo option (typically making no purchase). Under traditional CA, it would be unclear how to proceed with such data. However, RUT makes it clear that there is a considerable amount of preference information in this "pick-any" choice task on which to estimate a choice model due to several implied order comparisons. Every alternative chosen as preferred to the status quo effectively generates a separate choice set containing that alternative and the status quo; similarly, separate choice sets can be generated for each alternative that was not preferred to the status quo, where the status quo should be preferred. Next consider a case where one randomly assigns choice sets to people and none of the *k* offered alternatives in a choice set have attributes; each person is asked to choose their preferred alternative. Clearly a choice model can be estimated from the data in this task that specifies only alternative-specific constants. Further, one can ask respondents (in structured or free format styles) to indicate what attributes (and associated levels) they associate with each alternative, with which data can be used to estimate a choice model. Both examples yield SP data that bears little to no resemblance to traditional CA data, except insofar as there are stated responses and a "design matrix." However, unlike traditional CA data, these sources of preference information can be used to estimate discrete choice models consistent with RUT.

Another way to demonstrate the versatility of DCEs is to combine data from different DCE tasks or studies. Earlier, we noted that SP data from a DCE could be combined with RP data once the structure of both forms of data are clearly recognized (see equation 4). The same logic applies to combining data from DCEs with differing preference elicitation formats, ranges of alternatives offered to respondents, nature of

decision making contexts exemplified by the DCE, and samples used. The statistical framework for combining DCE data from different sources was provided by Hensher et al. (1999) and by Swait and Louviere (1993). The key insight is that if the vectors of model preference parameters are the same for all data sources but random component variances differ, vectors of estimated preference parameters should be proportional to one another. Several examples of this approach for the case of retail choices are in Severin et al. (2001).

Finally, the large body of work in economics on RUT is applicable to DCEs. Thus, researchers using DCEs have immediate access to consistent models to deal with complex issues like sample selection and pooling discrete choices and continuous responses (e.g., Greene 2007). DCEs also are anchored in utility maximization, so one can compare the effects on observed responses of different types of information and/or strategic incentives (e.g., Carson and Groves 2007) used in DCEs. In contrast, traditional CA should be seen for what it really is, namely a purely descriptive way to fit a statistical model to a set of observed ranking or rating data with no ability to inform questions about how consumer behavior is likely to change in response to changes in the choice context.

7 Conclusion

The terms discrete choice experiment (DCE) and conjoint analysis (CA) should not be considered synonymous, nor should DCEs be viewed as a special case of CA.[9] While both may use experimental designs from statistics and/or particular software packages may "suggest" similar methods, this is largely an illusion. Traditional CA is based on CM, while DCEs are based on RUT. RUT is a well-tested theory strongly associated with error components whose properties play key roles in parameter estimates and welfare measures derived from DCE data collection. In contrast, traditional CA is largely a curve fitting/scaling exercise where error components are largely ad hoc and lack clear interpretations. Moreover, by using DCEs to gain a better understanding of how people make choices, practitioners are likely to learn how to construct better empirical studies in marketing and other applied economic fields where choices play important roles.

Acknowledgements

We are grateful to Accenture (and Hikaru Phillips) for allowing us to draw extensively on an earlier Memetrics' White Paper "Why Stated Preference Discrete Choice Modelling is NOT Conjoint Analysis (and what SPDCM IS)" by Jordan Louviere (2000) for this paper.

References

Adelman, L., T. R. Stewart, and K. R. Hammond. 1975. A case history of the application of social judgment theory to policy formulation. *Policy Sciences,* 6 137-159.

[9] In contrast, CM, which underlies traditional CA, might be thought of as a special case of RUT which is routinely rejected by both RP and SP data.

Anderson, J. L., and S. U. Bettencourt. 1993. A Conjoint Approach to Model Product Preferences: The New England Market for Fresh and Frozen Salmon. *Marine Resource Economics,* 8 31-49.

Anderson, N. H. 1962. Application of an Additive Model to Impression Formation. *Science,* 138 (3542) 817-818.

Anderson, N. H. 1970. Functional Measurement and Psychophysical Judgement. *Psychological Review,* 77 (3) 153-170.

Anderson, N. H., and J. Shanteau. 1977. Weak Inference With Linear Models. *Psychological Bulletin,* 84 (6) 1155-1170.

Ben-Akiva, M., and T. Morikawa 1990. Estimation of Switching Models from Revealed Preferences and Stated Intentions. *Transportation Research,* 24A (6) 485-495.

Bradlow, E. T., Y. Hu, and T.-H. Ho. 2004. A Learning-Based Model for Imputing Missing Levels in Partial Conjoint Profiles. *Journal of Marketing Research,* 41 369-381.

Carson, R. T., and T. Groves. 2007. Incentive and informational properties of preference questions. *Environmental and Resource Economics,* 37 181-200.

Carson, R. T., and J. J. Louviere. Forthcoming. A Common Nomenclature For Stated Preference Elicitation Approaches. *Environmental and Resource Economics*.

Carson, R. T., J. J. Louviere, D. A. Anderson, P. Arabie, D. S. Bunch, D. A. Hensher, et al. 1994. Experimental Analysis of Choice. *Marketing Letters,* 5 (4) 351-368.

Cattin, P., and D. R. Wittink. 1982. Commercial use of conjoint analysis: A survey. *Journal of Marketing,* 46 44-53.

Chapman, R. G., and R. Staelin. 1982. Exploiting Rank Ordered Choice Set Data Within the Stochastic Utility Model. *Journal of Marketing Research,* 19 288-301.

Dawes, R. M., and B. Corrigan. 1974. Linear models in decision making. *Psychological Bulletin,* 81 (2) 95-106.

DeSarbo, W. S., M. Wedel, M. Vriens, and V. Ramaswamy. 1992. Latent Class Metric Conjoint Analysis. *Marketing Letters,* 3 (3) 273-288.

Fiebig, D.G., M. P. Keane, J. Louviere, and N. Wasi. 2010. The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity, *Marketing Science*, 29 (3) 393-421.

Flynn, T. N., J. J. Louviere, A. A. J. Marley, J. Coast, and T. J. Peters. 2008. Rescaling quality of life values from discrete choice experiments for use as QALYs: a cautionary tale. *Population Health Metrics,* 6 1-6.

Green, P. E. and V. R. Rao. 1971. Conjoint Measurement for Quantifying Judgmental Data, *Journal of Marketing Research*, 8, 355-363.

Green, P. E., A. M. Krieger, and Y. J. Wind. 2001. Thirty years of conjoint analysis: Reflections and prospects. *Interfaces,* 31 S56-S73.

Green, P. E., and V. Srinivasan. 1978. Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research,* 5 103-123.

Greene, W. H. 2007. Econometric analysis. Upper Saddle River: Prentice-Hall.

Harrison, R. W., J. Gillespie, and D. Fields. 2005. Analysis of Cardinal and Ordinal Assumptions in Conjoint Analysis. *Agricultural and Resource Economics Review*, 34 (2) 238-252.

Hensher, D. A., J. J. Louviere, and J. Swait. 1999. Combining sources of preference data. *Journal of Econometrics*, 89 197-221.

Hensher, D.A., J.M. Rose, and W.H. Greene. 2005. *Applied choice analysis: A primer*. Cambridge: Cambridge University Press.

Howard, J. A., and J. N. Sheth. 1969. The Theory of Buyer Behavior. New York: John Wiley & Sons.

Jedidi, K., and Z. J. Zhang. 2002. Augmenting Conjoint Analysis to Estimate Consumer Reservation Price. *Management Science*, 48 (10) 1350-1368.

Kahneman, D., and A. Tversky. 1979. Prospect theory: an analysis of decision under risk. *Econometrica*, 47 (2) 263-291.

Krantz, D. H., and A. Tversky. 1971. Conjoint measurement analysis of composition rules in psychology. *Psychological Review*, 78 (2) 151-169.

Louviere, J. J. 1988. Analyzing decision making: Metric conjoint analysis. Newbury Park: Sage Publications Inc.

Louviere, J. J., and T. Eagle. 2006. Confound it! That pesky little scale constant messes up our convenient assumptions. Sawtooth Conference: http://www.sawtoothsoftware.com/download/techpap/2006Proceedings.pdf pp. 211-228). Delray Beach, Florida, USA.

Louviere, J. J., D. A. Hensher, and J. Swait. 2000. Stated choice methods: analysis and application. Cambridge: Cambridge University Press.

Louviere, J. J., T. Islam, N. Wasi, D. Street, and L. Burgess. 2008. Designing discrete choice experiments: do optimal designs come at a price? *Journal of Consumer Research*, 35 360-375.

Louviere, J. J., R. J. Meyer, D. S. Bunch, R. T. Carson, B. Dellaert, W. M. Hanemann, et al. 1999. Combining Sources of Preference Data for Modeling Complex Decision Processes. *Marketing Letters*, 10 (3) 205-217.

McCabe, C., J. E. Brazier, P. Gilks, A. Tsuchiya, J. Roberts, A. O'Hagan, et al. 2006. Using rank data to estimate health state utility models. *Journal of Health Economics*, 25 418-431.

McFadden, D. 1986. The choice theory approach to market research. *Marketing Science*, 5 275-279.

McFadden, D., and K. Train. 2000. Mixed MNL Models For Discrete Response. *Journal of Applied Econometrics*, 15 (15) 447-470.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.) Frontiers in Econometrics. New York: Academic Press, 105-142.

Michell, J. 1990. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, N.J.: Erlbaum.

Moore, W. L., J. Gray-Lee, and J. J. Louviere. 1998. A Cross-Validity Comparison of Conjoint Analysis and Choice Models at Different Levels of Aggregation. *Marketing Letters*, 9 195-207.

Park, C. S. 2004. The Robustness of Hierarchical Bayes Conjoint Analysis Under Alternative Measurement Scales. *Journal of Business Research*, 57 1092-1097.

Severin, V., J. J. Louviere, and A. Finn. 2001. The stability of retail shopping choices over time and across countries. *Journal of Retailing*, 77 185-202.

Swait, J., and J. J. Louviere. 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research*, 30 305-314.

Thurstone, L. L. 1927. A law of comparative judgment. *Psychological Review*, 34 273-286.

Timmermans, H. J. P., R. E. C. M. van der Heijden, and H. Westerveld. 1982. The identification of factors influencing destination choice : an application of the repertory grid methodology. *Transportation*, 11 (2) 189-203.

Varian, H. (1992). Microeconomic Analysis. New York: Norton.

Vriens, M., H. R. van der Scheer, J. C. Hoekstra, and J. R. Bult. 1998. Conjoint Experiments for Direct Mail Response Optimization. *European Journal of Marketing*, 32 323-339.

Wittink, D. R., and P. Cattin. 1989. Commercial use of conjoint analysis: an update. *Journal of Marketing*, 53 91-96.

Yellott, J. I. 1977. The relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution. *Journal of Mathematical Psychology*, 15 109-144.