CONFOUND IT! THAT PESKY LITTLE SCALE CONSTANT MESSES UP OUR CONVENIENT ASSUMPTIONS

JORDAN LOUVIERE
UNIVERSITY OF TECHNOLOGY, SYDNEY
THOMAS EAGLE
EAGLE ANALYTICS, INC.

INTRODUCTION

Since the early 1990s there has been much progress in understanding and taking into account preference heterogeneity in probabilistic discrete choice models (e.g., Wedel and Kamakura 1999; McFadden and Train 2000). The vast majority of models applied in marketing and applied economics try to represent heterogeneity as some type of discrete or continuous distribution of preferences. These relatively new types of statistical models have done well in comparisons against simpler model forms like conditional multinomial logit in terms of in- and out-of-sample fits, with fit performance often assessed against so-called "hold-out" sets. It is fair to say that these models are long on statistical theory, but short on behavioral theory; the latter aspect is the focus of this paper.

In particular, scientists typically spend time a) formulating hypotheses and/or theory, b) testing deductions from theory and/or hypotheses, or c) testing assumptions underlying theory and/or hypotheses. The focus of this paper is on testing certain key assumptions about error distributions implicit in all limited dependent variable models. It is rare to see these assumptions tested, but it is important to do this to allow researchers to know if the assumptions are satisfied, and to discuss the consequences if they are not.

To preface our discussion and results, we note that all statistical models in which the dependent variable is latent confound estimates of model parameters with error variability. Thus, if error variances are not constant across individuals and choice sets, estimated model parameters will vary with differences in error variances. Moreover, if error variances are not constant one cannot estimate unconfounded discrete or continuous distributions of preferences, and the consequences of not satisfying this assumption can be serious. This paper focuses on the error variance (or "unobserved heterogeneity") associated with probabilistic discrete choice models. We first discuss the role of the error variance in simple and more complex choice models, then we discuss and illustrate the confound between model parameters and error variance, and then we present and review academic research that demonstrates that it is unlikely that error variances are constant, but instead it is much more likely that the error variance is systematically related to a number of factors that we outline and discuss. Then we propose and discuss ways to deal with systematic variation in error variances. We conclude with some cautions about making claims based on current models, and summarize the key points in the paper.

The Scale Parameter in MNL

The familiar MNL model is a random utility model in which the errors are assumed to be independent and identically distributed (iid) as Type 1 Extreme Value random variates. That is, recall the familiar axiom of random utility theory:

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

where U_{jn} is the latent utility that individual n associated with choice option j; V_{jn} is the systematic or mean utility that individual n associates with option j; and ε_{jn} is the random or stochastic component of the utility of option n for individual n, which is assumed to be distributed as Extreme Value Type 1. If individual n seeks to maximize her utility, we can model the probability that she chooses option j as follows:

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

where all terms are as previously defined, except for $P(j|C_n)$, which is the probability that option j is chosen by a randomly chosen individual n facing choice set C_n. If the errors are iid Extreme Value Type 1, expression (2) leads to a closed-form expression for the choice probabilities called the MNL (multinomial logit) choice model (McFadden 1974). The consequences of these error assumptions are that individuals are "preference clones" who share the same fixed set of preference parameters. Variability in choices arises due to analysts' misspecification of true utility functions, inability to account for all relevant factors in choice, and other omissions and commissions.

These assumptions imply that the main diagonal of the error variance-covariance matrix is constant, and all off-diagonal error covariances equal zero. The vector of unknown model parameters can be expressed as a generalized regression function:

$$V_{jn} = \Sigma_k \beta_k X_{kn} + \varepsilon_{jn}, \quad (3)$$

where all terms are as previously defined, except for β_k and X_{kn}. β_k is a vector of empirical parameters associated with a vector of factors that underlie choices, X_{kn}. The vector of parameters is subscripted only with respect to the factors ("attributes") because they are fixed for all individuals. The vector of factors is doubly subscripted to indicate that factors vary over k dimensions but also (potentially) over people. The confound between error variance and the estimated parameters can be expressed as follows:

$$V_{jn} = \Sigma_k \lambda \beta_k X_{kn} + \varepsilon_{jn}, \quad (3)$$

where all terms are as previously defined, except for λ, which is a "scale parameter". Formally, in the case of MNL, $\lambda = SQRT(\pi^2 / 6 \sigma_\varepsilon^2)$, where π is the natural constant 2.141…, and σ_ε is the standard deviation of the error distribution. We say that λ "scales" the vector of parameters because each parameter, β_k, actually is $\beta_k/\sigma_\varepsilon$.

IMPLICATIONS OF SCALE IN MNL

The confound of scale and model parameters creates a fundamental identification problem, with the consequence that MNL model parameters cannot be identified unless λ is fixed to some constant (almost always 1.0). Thus, the parameters "are identified up to scale", which means that they can be identified once a constant is selected. The confound has no impact on predicted probabilities in MNL if the error assumptions are satisfied. This confound is not new, and discussions can be found in many sources like Ben-Akiva and Lerman (1985), Swait and Louviere (1993) and Louviere, Hensher and Swait (2000). What is new is that more researchers now realize that it is unlikely that error variances (and hence, scale) are constant in empirical data; instead, it is more likely that error variances systematically vary with attribute levels varied in choice experiments and real markets, as well as differences in individuals. For example, we later show that error variances systematically vary with levels of attributes, and there is evidence that survey respondents with low literacy skills have higher error variances than those with higher skills (See Louviere, Hensher and Swait 2000, Chapter 13).

A consequence of the confound is that it impacts magnitudes of estimated model parameters, and by implication, statistical inference. Specifically, smaller error variances (large scales) lead to larger model parameters, while larger error variances (small scales) lead to smaller model parameters, all else equal. Not surprisingly, standard errors of parameter estimates also are impacted - smaller scales lead to less precise estimates. Discussions of such issues can be found in workshop reports published in Marketing Letters since the advent of the modern Invitational Choice Symposia that were proposed and initially organized by one of the present authors (1989, Banff, Alberta, Canada). For example, discussion of the scale parameter ands its implications are in Louviere, et al, (1999), Louviere, et al (2002) and Louviere, et al (2006). We now turn our attention to the consequences if error variances are not constant.

WHAT IF SCALE IS NOT CONSTANT?

It is surprising that researchers have largely focused on preference heterogeneity to the exclusion of most other likely sources of unobserved variability. There has been little research into variance component models for discrete choices (for an exception, see Cardell 1997) that explicitly recognize that errors can be decomposed into systematic components associated with differences within and between individuals (the latter can be preference heterogeneity, but may be related to other factors that differ between individuals), environmental and context differences, temporal and spatial differences and other sources. Louviere, et al (1999) discuss these issues in detail, so we only briefly summarize them here.

They suggest that scale is impacted by many factors that can be summarized in the following general expression:

$$Y \mid X, Z, C, G, T, \qquad (4)$$

where Y = behavioral outcomes of interest; X = directly observable or manipulated variables; Z = characteristics of the individuals; C = factors that vary over conditions, contexts, circumstances, or situations; G = geographical, spatial or environmental factors that are relatively constant in one place, but may vary from place to place; and T = time varying factors. Thus, it is highly unlikely that scale is constant; it is much more likely to be systematically impacted by a wide array of factors. Also, error variability is unlikely to be unidimensional, but probably varies a) within consumers; b) between consumers; c) with measurement instruments; d) with market and environmental differences; and e) with many other sources.

Any one data source confounds many of the above sources of error variability, as discussed by Louviere, et al (1999, 2002). So, most researchers have samples of size = 1, but claim meaningful results! Single data sources limit generalizeability in many cases where potential sources of – say – temporal and spatial variability are constant, making it unclear how to generalize to past or future time periods or spatial locations. Thus, much more research into combining sources of preference and choice data is needed.

SCALE OR PREFERENCE HETEROGENEITY?

During the past decade many published choice models assumed that individuals are heterogeneous in their preferences. Very few publications suggested that individuals also might be heterogeneous in their error variances (or scales). We earlier noted that model parameter estimates and scale are confounded; hence, if scale varies across individuals, distributions of preference parameters will be confounded with distributions of scales. Later we provide evidence that this occurs in empirical data, but we note that because model estimates actually are $\beta_k/\sigma_\varepsilon$, it should be clear that one can have a distribution of true β_k if and only if σ_ε is constant. If β_k and σ_ε vary across individuals, observations, contexts, time and space, one cannot estimate a distribution of β_k without separating β_k and σ_ε. Of course, one can try to capture these effects by estimating higher moments of assumed distributions, but adding more statistical complexity in the form of additional latent effects does not seem warranted in the absence of better behavioral theory.

That is, little behavioral theory is evident in recent choice modeling papers; most authors rely on statistical theory. Limitations of single data sources suggest that without theory to suggest how components of variance differ by individuals, markets, contexts, experiments, etc, adding higher moments to choice models is probably a bad idea unless these distributions are constant over such sources of variation. One only needs to consider predicting infrastructure projects like toll roads or bridges years into the future to see that estimating higher moments is a bad idea as they also have to be forecast into the future. A better way forward is to develop theory and methods to capture variability differences. We now show how easy it is to confuse heterogeneity in model parameters and scale, which should give researchers reasons to think before estimating higher moments.

Consider two cases involving ten people in Table 5: 1) everyone has identical preferences for travel times and costs of public bus systems, with scales equal to 1.0; only intercepts differ; 2) scales vary across people, holding everything else constant. A key takeaway is that if scales vary across people, they will seem heterogeneous in preferences even if they differ only in scale. The last two columns show the processes are equivalent.

Table 5: Consequences of Scale Varying Across Individuals

	Condition 1: Only intercepts vary					**Condition 2: intercepts & scale vary**					
Person	Inter 1	time1	cost1	scale1	u1	Inter 2	time2	cost2	scale2	u2	u1*scale2
0	-1.00	-1.5	-1	1	1.50	-0.20	-0.30	-0.20	0.20	0.30	0.30
1	-0.75	-1.5	-1	1	1.75	-0.60	-1.20	-0.80	0.80	1.40	1.40
2	-0.50	-1.5	-1	1	2.00	-0.70	-2.10	-1.40	1.40	2.80	2.80
3	-0.25	-1.5	-1	1	2.25	-0.25	-1.50	-1.00	1.00	2.25	2.25
4	0.00	-1.5	-1	1	2.50	0.00	-3.00	-2.00	2.00	5.00	5.00
5	0.25	-1.5	-1	1	2.75	0.15	-0.90	-0.60	0.60	1.65	1.65
6	0.50	-1.5	-1	1	3.00	0.20	-0.60	-0.40	0.40	1.20	1.20
7	0.75	-1.5	-1	1	3.25	1.35	-2.70	-1.80	1.80	5.85	5.85
8	1.00	-1.5	-1	1	3.50	1.20	-1.80	-1.20	1.20	4.20	4.20
9	1.25	-1.5	-1	1	3.75	2.00	-2.40	-1.60	1.60	6.00	6.00

The above example assumes that time1 = -1 and cost1 = -1

That is, the only difference in the left-hand and right-hand side parameters is scale. That is, if one multiplies the left-hand side parameters by scale2, one obtains the right-hand side parameters. The last two columns show that multiplying u1 (overall utility for a bus described by time = -1 and fare = -1 in condition 1) by scale2 produces the same outcome as u2 (overall utility for a bus described by a time = -1 and a fare = -1 in condition 2).

Differences in scales across people can be consequential, as we now show. For example, let person 0 from Table 5 face a choice among three buses as shown below:

Bus #	**Time**	**Cost**	**Utility from left-hand side**	**Choice Prob**	**Utility from right-hand side**	**Choice Prob**
1	-1	+1	-0.5	0.506	-0.1	0.367
2	+1	-1	-1.5	0.186	-0.3	0.301
3	0	0	-1.0	0.307	-0.2	0.332

If person 0 significantly differs from a scale of 1.0, which in this case is a scale of 0.2 (a smaller scale, or larger error variance), choice probabilities can differ a lot even though the person's preferences for time and cost do not change. Thus, it is not possible to make meaningful statements about preference parameter distributions without taking scale differences into account at the same time.

How and Why We Know That Variances (Scales) Differ

There has been a fair amount of research on scale. For example, reviews of research in this area can be found in Louviere, Hensher and Swait (2000), Louviere, et al (2002), Louviere, et al (2006); empirical work on size and scopes of error assumption violations can be found in Train and Weeks (2005), Louviere and Islam (2004), Louviere (2004), DeShazo and Fermo (2002), Swait and Adamowicz (2001a,b), Louviere, et al (2002), Dellaert, Brazell and Louviere (1999) and Ohler, et al (2000), to name only a few. The preceding discussion and cited sources discuss several consequences if error variances are not constant, such as:

1. Estimates of price sensitivities (elasticities) or other policy effects may be misleading;
2. Willingness-to-pay (WTP) or other policy estimates may be misleading.
3. Forecasts may be biased and may depart significantly from reality.
4. Randomness in parameters confounds scale and real preference heterogeneity.
5. Hyperparameters in HB and/or MIXL are affected, and Latent Class model differences may be misleading.

In particular, differences in people may be reflected in parameter estimates, scale differences and/or some combination of both. The table below shows that individuals can be in the same part of the probability distribution with small parameters and a large scale or small scale and large parameters. These outcomes are observationally equivalent, which implies that more research is needed to separate parameters and scale. The table also suggests that researchers need to recognize that scale can be related to factors varied in choice experiments, factors that managers control, individual differences, environment, context, temporal and/or spatial differences. We say this again to emphasize that ignoring variability sources and/or assuming them away is dangerous and limits generalizeability. It also begs the question of how to deal with scale; later we discuss two ways to deal with scale in choice experiments.

	True Preference Parameters	
Scale	**Small**	**Large**
Small	Complimentary Processes	Potentially Observationally Equivalent
Large	Potentially Observationally Equivalent	Complimentary Processes

As noted by several authors, such as Louviere, et al (1999), one must combine multiple sources of data to make real progress in separating components of variance. This is not merely an academic issue because one cannot correctly predict choices in real markets if variance components differ between model estimation data sources and the market(s) to be predicted. So, instead of being seen as a convenient (or annoying) statistical assumption, researchers should begin to recognize that scale (or error variance) is a behavioral phenomenon that needs to be understood in its own right.

We now discuss how scale varies with choice experiment design. Example one is from an honours thesis by Chelsea Wise (Louviere and Wise 2004). She designed four conditions to vary attributes: 1) only brand and price, 2) brand, price + four less salient attributes, 3) brand, price + four more salient attributes; and 4) brand, price + all attributes. Figure 1 graphs the relative variability from each of these conditions. The least variability occurs for the least complex task (only brand and price), followed by a task with brand, price and more salient attributes. Higher variability occurred for tasks with less salient attributes and all attributes. These results are not "surprising". It is surprising that inequality of variances in choice data has received so little research attention.

Figure 1: Relative Variance Differences in Experimental Conditions

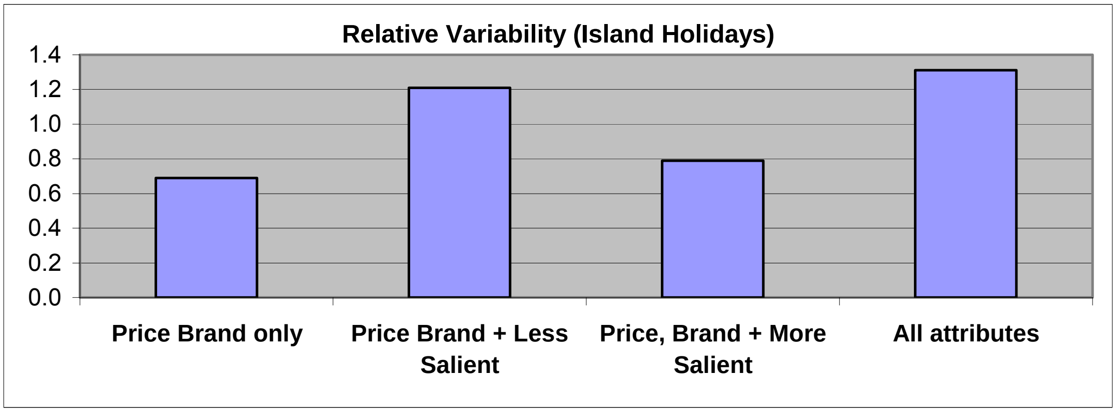


Figure 2 is the conditional (mean) price response curves estimated from Wise's four conditions, which show that price slopes differ systematically with relative variability in each condition. Thus, the four conditions exhibit different choice probabilities, and the differences are due to differences in choice variability.

Figure 2: Differences in Price Response Curves by Experimental Condition

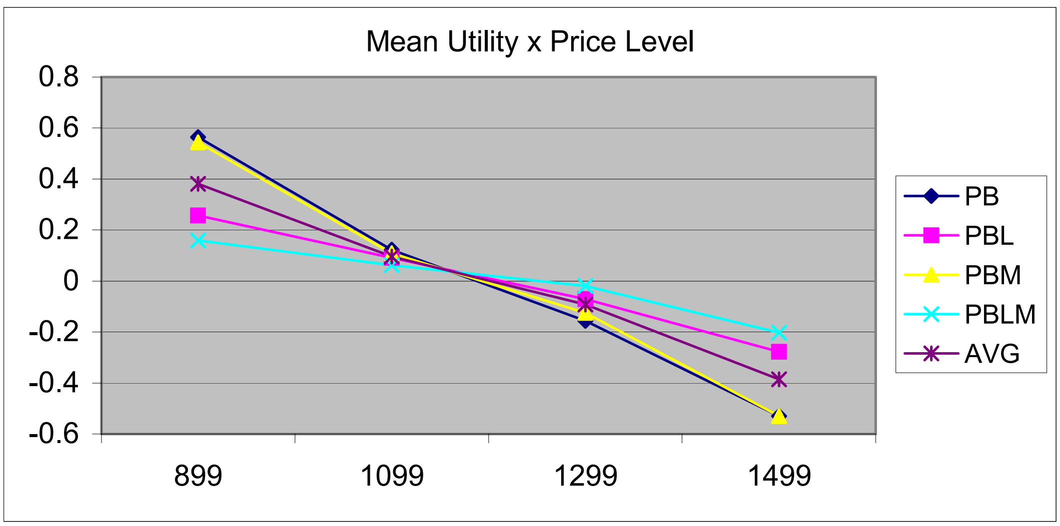
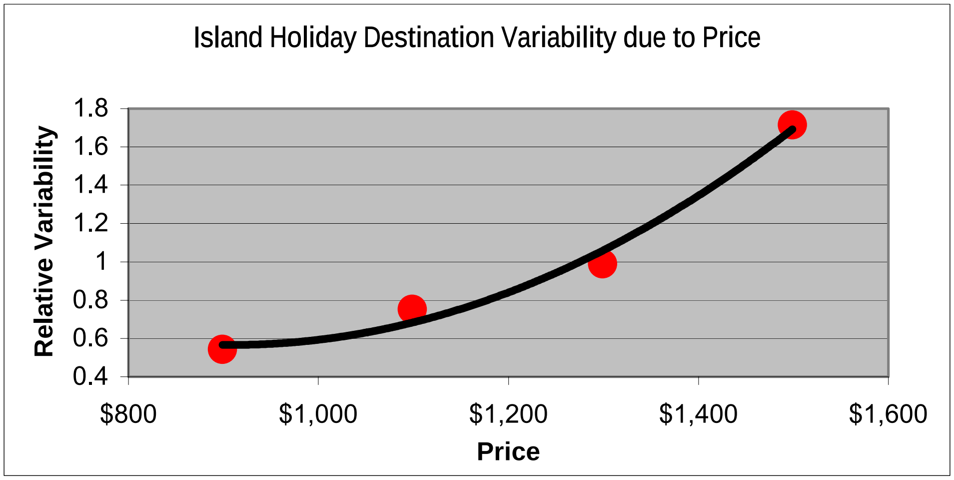


Figure 3 shows a relationship between relative variability and price levels in Wise's data. Each point is an estimate of the variability for each price level, analogous to a random coefficients result with a different standard deviation for each price level. The graph shows choice variability is not constant, but varies systematically with price levels.

Figure 3: Relationship between Relative Choice Variability and Price Levels



Islam and Louviere (2004) provide another example. They designed 64 experiments to vary attribute presence/absence. Brand and price were always present; the design dictated other attributes that vary. Figure 4 shows relative choice variability associated with each level of percent real fruit juice or the juice price. In the case of percent juice, there also is an estimate of the effect of this attribute being "missing".

Figure 4: Conditional Means for Percent Juice and Associated Variability

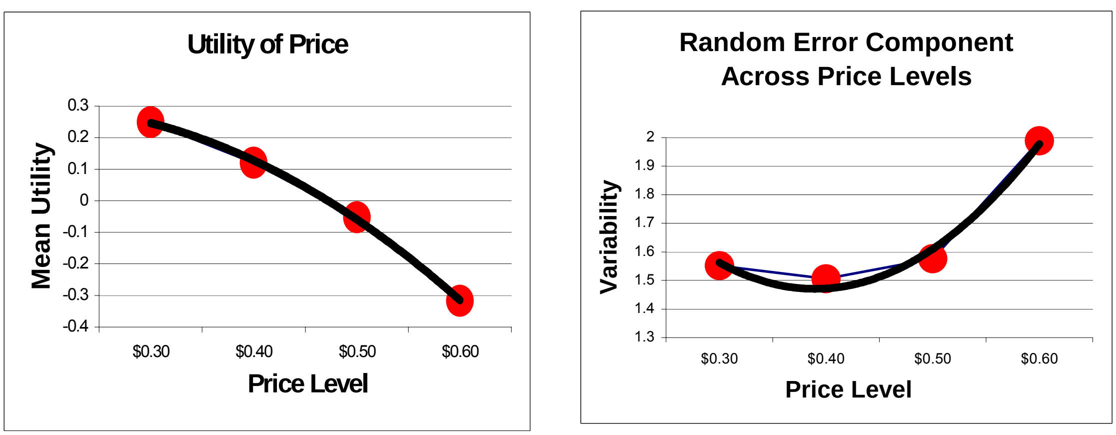
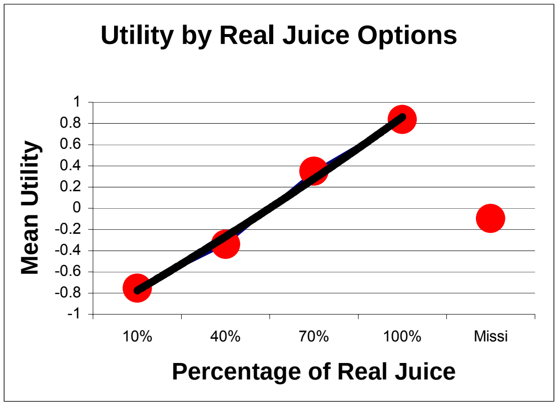
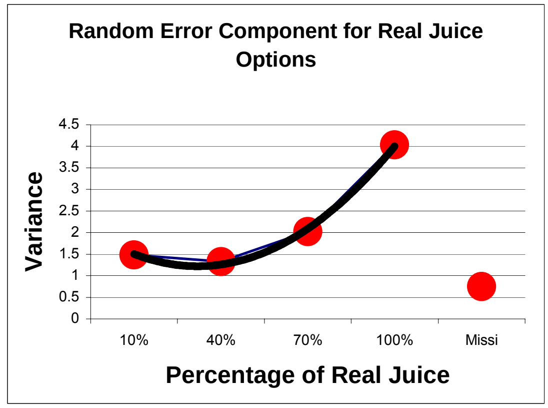


Figure 4 shows that the relative variance is an inverse-U shaped function of the percent real juice and price levels. Estimated effect of "missing" for percent real juice is nearly zero, implying that when absent, it doesn't affect choices. Also, when absent, the relative variance decreases. The relative variance increases when attributes are present, implying choice variability increases/decreases as one "adds"/"subtracts" more attributes.





We cited literature and reviewed empirical evidence that error variability in choice experiments is not constant but varies systematically with many factors. Now we discuss two ways that one can try to capture and control this variability in models.

CAN WE MODEL SCALE?

To model scale, one must specify scale as a function of observables. "Observables" are factors that can be identified and measured (preferably also forecast into the future over space, contexts, etc). For example, embedding attitudes in a model might "explain" some individual differences, but it is hard to forecast them into the future. So, including them in models requires strong assumptions about invariance over people, space, time, contexts, etc. This assumption is unlikely to hold, so this is not a particularly good idea.

Covariance Heterogeneity Models (CHMs) are growing in popularity in applied economics and marketing; examples include Swait and Adamowicz (1997; 2001a,b); Hensher, Louviere and Swait (1999); DeShazo and Fermo (2002) and Delleart, Brazell and Louviere (1999). Attractive CHM properties include: a) simple and interpretable, b) closed form, c) captures attribute interactions, even if not specified in mean components, and d) captures nonlinear effects even if main effects are linear. CHMs mimic random coefficient models if one models error variability and conditional response means jointly as a function of attributes varied in experiments. We can express this CHM as follows:

$$P(i|C) = \{\exp[\exp(\alpha_0 + \alpha_1 X_i)(\beta_0 + \beta_1 X_i)]\} \,/\, \Sigma_{j\in C}\, \{\exp[\exp(\alpha_0 + \alpha_1 X_j)(\beta_0 + \beta_1 X_j)]\}, \quad (5)$$

where α_0 is an intercept for the scale function, constrained to be positive (scale must be > 0); α_1 is a vector of parameters associated with a design matrix of attribute effects, X_i and X_j; β_0 is an intercept for the mean (systematic) function; $\beta_1 X_i$) is a vector of parameters associated with a design matrix of attribute effects, X_i and X_j.

We used CHMs to analyze 44 experiments designed by the UTS Centre for the Study of Choice (funded by the Australian Research Council). Experiments combined attributes (4, 8, 12, 16), numbers of attribute differences (nested under attributes), and relative design efficiency (37% to 100%). Subjects were recruited from an opt-in panel, with 100 randomly assigned to each experiment. The mean component was specified as a function of the effects-coded factors; the scale component as a function of logarithms of numbers of attributes and design efficiency. Estimation results are in Table 6.

Table 6: Results for CHMs Estimated from Pizza and Island Holiday Choices

Island Holiday Choices				Delivered Pizza Choices			
Effects	Estimate	T-Stat	P(T)	Effects	Estimate	T-Stat	P(T)
Price	-0.082	-9.890	0.000	Type of Pizza	-0.002	-0.330	0.740
Destination type	0.035	7.790	0.000	Price	-0.104	-7.180	0.000
Airline	-0.008	-2.500	0.013	Quality	-0.177	-7.380	0.000
Length of stay	0.126	10.590	0.000	Del time	-0.036	-6.040	0.000
Meals included	-0.125	-10.540	0.000	Crust	0.032	5.080	0.000
Tours available	-0.048	-8.800	0.000	Sizes	-0.053	-6.390	0.000
Season	0.000	0.110	0.909	Temp	-0.144	-7.410	0.000
Hotel type	-0.199	-10.790	0.000	Open hours	-0.023	-4.450	0.000
Length of Trip	-0.011	-2.890	0.004	Del Charge	-0.093	-6.870	0.000
Cultural activities	-0.037	-7.210	0.000	Store Type	0.064	6.250	0.000
Dist to attractions	-0.049	-7.920	0.000	Baking Method	-0.040	-4.830	0.000
Swimming Pool	-0.052	-8.140	0.000	Manners	-0.011	-1.620	0.105
Helpfulness	-0.026	-4.300	0.000	Vegetarian option	-0.064	-5.150	0.000
Type of Holiday	-0.045	-6.670	0.000	Delivery Guarantee	-0.056	-4.980	0.000
Beach availability	-0.068	-7.990	0.000	Distance	-0.041	-3.820	0.000
Brand	-0.001	-0.090	0.927	Variety	-0.039	-3.420	0.001
Intercept (8 atts)	0.002	0.230	0.815	Intercept (8 atts)	0.034	2.740	0.006
Intercept (12 atts)	-0.020	-1.950	0.051	Intercept (12 atts)	-0.014	-0.990	0.324
Intercept (16 atts)	0.029	3.290	0.001	Intercept (16 atts)	-0.040	-2.770	0.006
Scale Intercept	3.377	22.490	0.000	Scale_Intercept	4.437	23.600	0.000
Log (No. of Attrib.)	-0.571	-14.630	0.000	Log (No. of Attrib.)	-0.908	-23.390	0.000
Log (Efficiency/10)	-0.715	-20.230	0.000	Log (Efficiency/10)	-0.878	-18.250	0.000

The key takeaway in Table 6 is scale varies systematically with the number of attributes and design efficiency. Prior to estimating the model in Table 6, we graphed estimated scale against attributes and efficiency; this suggested both were approximately logarithmically related to scale. Efficient designs maximize attribute differences; so the results imply that more efficient designs that vary more attributes will increase choice variability at a decreasing rate, all else equal. The example shows that CHMs can capture systematic relationships between unobserved variability and factors varying within and between experiments, across individuals, contexts, etc. Now we discuss another way to capture unobserved variability, namely estimating choice models for single individuals.

CAN WE ESTIMATE CHOICE MODELS FOR INDIVIDUALS?

The Australian Research Council funded a team in the Centre for the Study of Choice at UTS to develop theory and design methods to estimate choice models for single people (Louviere, Marley, Street and Burgess 2004). We now briefly describe some empirical work that focused on potential constraints on sizes of problems that can be handled with this approach. Specifically, we designed 66 experiments to vary a) 11 combinations of two- and four-level attributes ranging from a $2^3 \times 4^3$ to a $4^8 \times 2^4$, b) number of options per choice set (3, 4 or 5) and c) category (delivered pizza or cross country flights). Subjects were web panelists recruited from an Australian opt-in panel; approximately 20 subjects were randomly assigned to each of the 66 conditions.

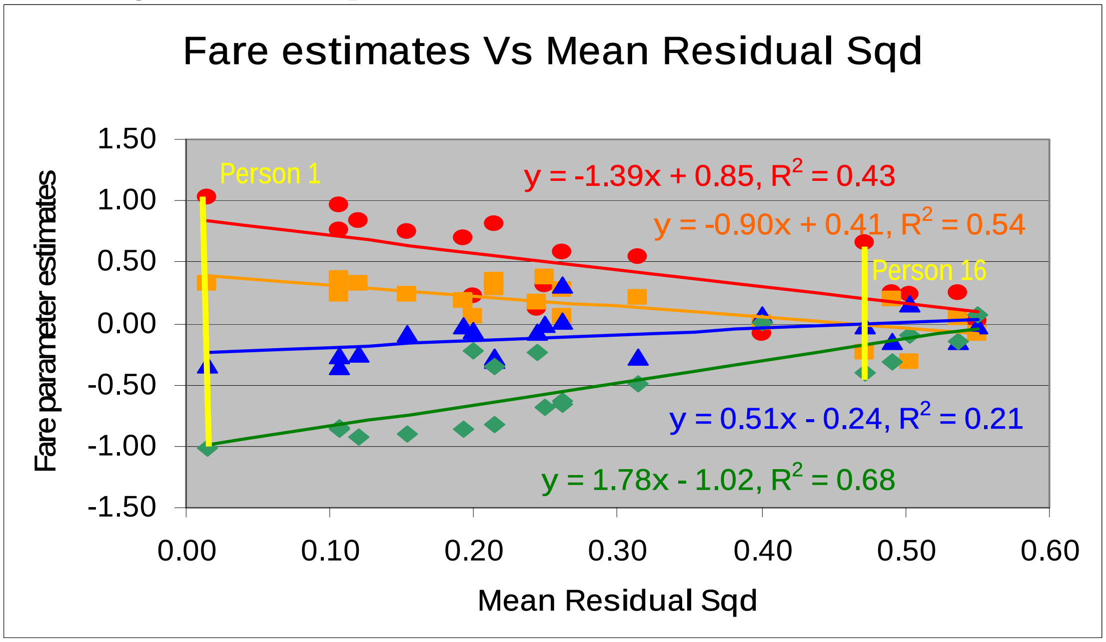
Below we report the results of one condition with 32 choice scenarios (sets), and four options described by 10 attributes ($3^3 \times 2^7$). To date we have analyzed the smallest and largest of the 66 conditions, and have been able to estimate models for all individuals in these conditions. This one condition is one of the larger conditions, but otherwise is not unique in any way; the model estimation results are summarized in Table 8, which can be interpreted like the results of a random coefficient model. Table 8 contains the summary statistics for effects-coded MNL models for the individuals in the sample. One does not need to make assumptions about distributions of preference parameters in this case because (by definition) one has the empirical distribution for this sample population. A key takeaway from Table 8 is that standard errors for numerical attributes like fare are not constant; they are systematically related to the attribute levels.

Table 8: Summary Statistics for Individual Models in One Condition

Effect	N	Mean	StdErr	StdDev	T-Stat
ASC1	20	0.0205	0.0511	0.2286	0.4005
ASC2	20	0.0409	0.0530	0.2369	0.7718
ASC3	20	-0.0134	0.0500	0.2237	-0.2675
Flying time1	20	0.1507	0.0562	0.2511	2.6845
Flying time2	20	0.0306	0.0236	0.1057	1.2925
Flying time3	20	0.0115	0.0187	0.0834	0.6173
Fare1	20	0.4650	0.0750	0.3352	6.2037
Fare2	20	0.1550	0.0434	0.1943	3.5678
Fare3	20	-0.1013	0.0392	0.1753	-2.5842
Checkin	20	0.0226	0.0179	0.0801	1.2597
Airline1	20	0.1263	0.0450	0.2010	2.8103
Airline2	20	-0.0289	0.0335	0.1500	-0.8633
Airline3	20	-0.0175	0.0399	0.1786	-0.4395
Meals	20	-0.0423	0.0127	0.0570	-3.3222
Entertainment	20	-0.0151	0.0097	0.0434	-1.5535
Wait time for Bags	20	0.0678	0.0293	0.1312	2.3119
Frq Flyer rewards	20	-0.0059	0.0165	0.0737	-0.3584
Number of Stops	20	0.0153	0.0092	0.0410	1.6686
%OnTime departures	20	0.0442	0.0185	0.0826	2.3945
Free Alcohol	20	0.0138	0.0145	0.0648	0.9511
Free Drinks	20	-0.0629	0.0174	0.0776	-3.6273

Experimental subjects were asked to choose most and least preferred options in each scenario; questions were repeated to order options, which also yields frequencies of choices in each scenario. This allows calculation of individual error variances (sums of squared residuals). Figure 6 is a graph of mean squared residual sums (MRSs) versus four fare level estimates across individuals. Random utility theory predicts four straight lines converging to zero as the MRS increase. Each person is represented by four points on the graph; vertical lines on left- and right-hand sides indicate two subjects.

Figure 6: Mean Square Residuals Versus Fare Parameter Estimates

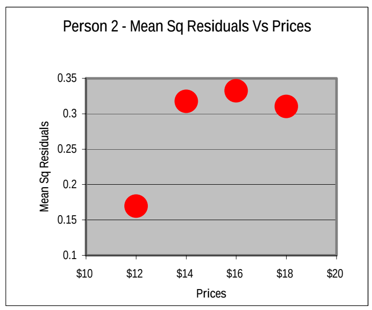


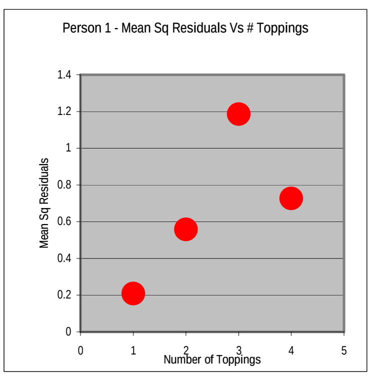
A key takeaway in Figure 6 is that variation in estimated fare parameters is due largely to MRS differences. That is, Figure 6 implies that a "random scale" model (a distribution of individual scales) should fit as well as (if not better than) a fare preference distribution model. We can test this hypothesis by interacting the effects-coded design columns associated with each four level attribute with the individual MRS values. If the MRS values contribute nothing to the model fit, all interactions should be non-significant.

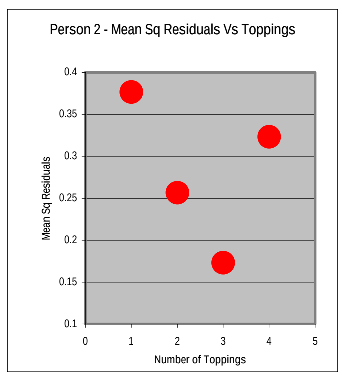
Table 8: Testing MRS by Attribute Level Effects

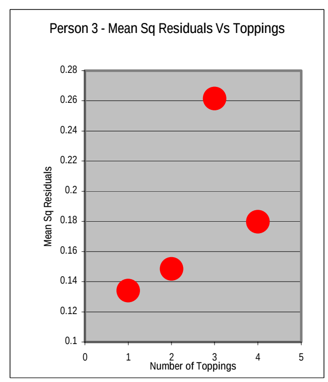
Effect	Estimate	StdErr	T-Stat	P(T)
asc1	-0.212	0.050	-4.256	0.000
asc2	-0.151	0.049	-3.061	0.002
asc3	-0.127	0.050	-2.529	0.011
asc1 x MRS	0.912	0.138	6.612	0.000
asc2 x MRS	0.753	0.138	5.462	0.000
asc3 x MRS	0.447	0.145	3.082	0.002
Flying time1	0.286	0.037	7.805	0.000
Flying time2	0.032	0.039	0.838	0.402
Flying time3	-0.062	0.039	-1.609	0.108
FT1 x MRS	-0.475	0.117	-4.062	0.000
FT2 x MRS	-0.091	0.120	-0.760	0.447
FT3 x MRS	0.169	0.120	1.411	0.158
Fare1	0.768	0.034	22.490	0.000
Fare2	0.319	0.039	8.166	0.000
Fare3	-0.198	0.044	-4.480	0.000
Fare1 x MRS	-1.247	0.108	-11.576	0.000
Fare2 x MRS	-0.646	0.123	-5.267	0.000
Fare3 x MRS	0.373	0.132	2.831	0.005
Airline1	0.058	0.039	1.479	0.139
Airline2	-0.059	0.038	-1.545	0.122
Airline3	0.020	0.011	1.861	0.063
Aline1 x MRS	0.330	0.116	2.846	0.004
Aline2 x MRS	-0.293	0.124	-2.365	0.018
Aline3 x MRS	0.178	0.119	1.494	0.135
Meals	-0.032	0.011	-3.011	0.003
Entertain	-0.015	0.011	-1.364	0.173
GetBags	0.054	0.011	4.948	0.000
FrqFlyer	-0.006	0.011	-0.551	0.582
#Stops	0.012	0.011	1.042	0.298
%OnTime	0.036	0.011	3.325	0.001
Alcohol	0.009	0.011	0.791	0.429
Drinks	-0.058	0.011	-5.311	0.000
Checkin	0.029	0.038	0.767	0.443

MRS x attribute level interactions are bolded in Table 8. Many interactions are highly significant; hence, individual-level error variance differences significantly impact model results. The results also suggest that individuals differ in levels of error variability; hence, constant error variability assumptions are wrong; the results also are not unique to this experimental condition. Constant variances within individuals also are unlikely. The graphs below show that individual-level MRSs values are systematically related to levels of prices and pizza toppings for the first three individuals in the Table 8 dataset. Again, there is nothing unusual about these three individuals; we chose only three to save space. The graphs suggest that error variances are systematically related to the attribute levels for each person. Thus, error variances are not constant within individuals, either.





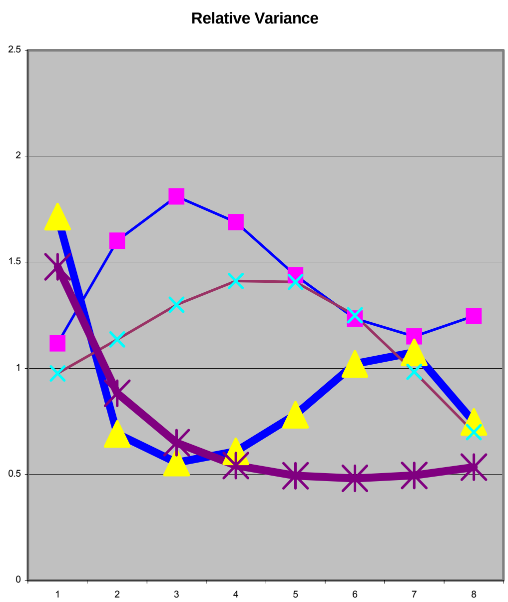


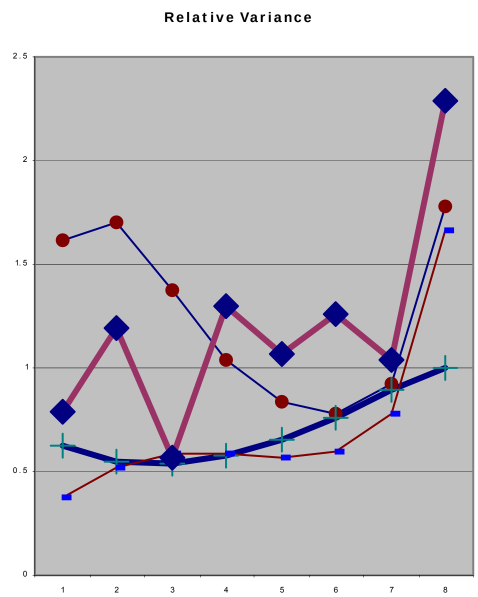


Does Error Variability Differ by Scenario Order?

A team from the UTS Centre for the Study of Choice helped design a UK experiment to test several hypotheses about welfare measures and choice experiments (led by Ian Bateman, currently editor of EARE). We tested order effects on choice variability by designing a target 2^3 factorial that described changes in water quality and costs relative to a status quo (other designed scenarios are not germane to the test). The 8 targets were shown as the first or last eight scenarios; we also varied whether subjects saw/did not see a glossary with all attributes and levels before the scenarios and whether a first (non-target) scenario described a large or small change in quality. We used a latin square to control for order, which created 8 more conditions that allow us to estimate models for each order condition. Individuals were randomly assigned to a particular order condition x first/last condition x glossary x quality change.

We estimated CHMs from the data for the glossary x quality change conditions for the first and second set of eight scenarios, allowing separate scale parameters for each order. Below we graph these results for each condition. Scenario order is on the X-axis; points on the graph are estimates of scale for each of the four conditions. The left-hand side graph represents the condition where the target scenarios appeared as the first 8; the right-hand graph represents the condition where the target appeared as the second eight. These graphs show that variance systematically varies across scenario orders.





DISCUSSION AND CONCLUSIONS

All latent choice models confound scale and parameter estimates. The confound poses particular problems in complex models like random coefficients models and latent class models. If one merely wants to predict choice probabilities from choice models estimated from experiments (seemingly the majority of applications), the models will predict well but will be biased and incorrect. That is, it is likely that a) many random coefficient models are over-fit, b) error distribution assumptions are not satisfied, and c) as Train and Weeks (2005) note, some distribution combinations make little theoretical sense.

We began by noting the scientists care about assumptions that underlie models, and this paper was about such concerns. That is, we were concerned about assuming that errors are iid, a widespread assumption in random coefficient models. We noted that model parameter estimates are "scaled" by standard deviations of error distribution, and unless this standard deviation is constant for all observations, distributions of model parameter estimates will be confounded with distributions of error variances. Evidence was provided to show that it is very unlikely that errors are constant; instead, they are likely to be systematically related to different factors that we noted in the paper.

So, the bottom line is that one cannot estimate individual-level parameters from complex choice models unless one can separate scale and model parameter estimates. We discussed two potential ways to do this: using various forms of covariance heterogeneity models and developing ways to estimate models for single persons. We illustrated both with empirical results that reveal systematic relationships between the attribute levels manipulated in choice experiments and error variability. Thus, it is highly likely that random coefficient models are biased and misleading. Moreover, it is well-known that virtually all choice models fit to choice experiment data will fit the data well (See, e.g., Dawes and Corrigan 1974), and so good fits and predictions to sets of hold-out choices are largely useless to test the validity of choice models estimated from experiments.

The field needs research that will lead to new models that can capture both scale and systematic component (mean) effects. The field also would benefit from research that leads to better and more useful behavioral theory. What the field definitely does not need is more complex statistical models, and it would be beneficial for academics in marketing to admit that there are serious issues associated with these classes of models, and that just because one can formulate and estimate complex statistical models does not mean that one should in fact do this. So, it is now time for marketing researchers to acknowledge these issues and to move on to more promising and less obviously empirically incorrect methods and models.

References Cited

Ben-Akiva, M. and Lerman, S.R. (1985) Discrete Choice Analysis. Cambridge: MIT Press.

Cardell, N.S. (1997) "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," Econometric Theory, 13, 185-213.

Dellaert, B., Brazell, J. and Louviere, J.J. (1999) "The Effect of Attribute Variation on Consumer Choice Consistency," Marketing Letters 10(2), 139-147.

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(1), 123-143.

Hensher, D., Louviere, J.J., Swait, J. (1999) "Combining Sources of Preference Data," Journal of Econometrics, 89, 197-221.

Louviere, J.J. (2001) "What if Consumer Experiments Impact Variances as Well as Means? Response Variability as a Behavioral Phenomenon," Journal of Consumer Research, 28, 506-511.

Louviere, J.J. (2004) "Complex Statistical Choice Models: Are the Assumptions True, and If Not, What Are the Consequences?" CenSoC Working Paper No. 04-002, Centre for the Study of Choice, University of Technology, Sydney, http://www.business.uts.edu.au/censoc/papers/index.html

Louviere, J.J. and Islam, T. (2004) "To Include or Exclude Attributes in Choice Experiments: A Systematic Investigation of the Empirical Consequences," Wiley, J & Thirkell, P. (eds.), Proceedings of the Australian and New Zealand Marketing Academy Conference, Victoria University of Wellington, Wellington, New Zealand, 2004.

Louviere, J. J., Hensher, D. A. and Swait, J. (2000) Stated Choice Methods: Analysis and Applications, Cambridge University Press.

Louviere, J.J., Burgess, L., Street, D. and A.A.J. Marley (2004), "Modeling the Choice of Single Individuals By Combining Efficient Choice Experiment Designs with Extra Preference Information," CenSoC Working Paper No. 04-005, , Centre for the Study of Choice, University of Technology, Sydney, http://www.business.uts.edu.au/censoc/papers/index.html

Louviere, J.J., Street, D., Carson, R., Ainslie, A., DeShazo, J.R., Cameron, T., Hensher, D., Kohn, R. and A.A.T. Marley (2002) “Dissecting the Random Component of Utility,” Marketing Letters, 13, 3, 177-193.

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

Louviere, J., Train, K., Ben-Akiva, M., Bhat, C., Brownstone, D., Cameron, T.A., Carson, R.T., DeShazo, J.R., Fiebig, D., Greene, W., Hensher, D. and Waldman D. (2006) “Recent Progress on Endogeneity in Choice Modelling,” Marketing Letters, 16, 3-4.

McFadden, D. (1974) “Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.), Frontiers in Econometrics, 105-142, Academic Press: New York.

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

Ohler, T., Le, A., Louviere, J.J. and J. Swait (2000) “Attribute Range Effects in Binary Response Tasks,” Marketing Letters, 11, 3 (August), 249-260.

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

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

Swait, J. and Louviere, J.J. (1993) “The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models,” Journal of Marketing Research, 30, 305-314.

Train, K. and Weeks, M. (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, Springer Publishers: Dordrecht, The Netherlands, Chapter 1, pp. 1-17,

Wedel, M. and Kamakura, W. (1999) Market Segmentation: Conceptual and Methodological Foundations, Dordrecht: Kluwer Academic Publishers.

Wise, C., Louviere, J.J. (2004) “The Impact of Varying Amounts of More and Less Salient Product Information Upon Consumer Willingness-to-Pay.” In Wiley, J. & Thirkell, P. (eds.), Proceedings of the Australian and New Zealand Marketing Academy Conference, Victoria University of Wellington, Wellington, New Zealand.