MAKING MAXDIFF MORE INFORMATIVE: STATISTICAL DATA FUSION BY WAY OF LATENT VARIABLE MODELING

LYND BACON
YOUGOV/POLIMETRIX, INC.
PETER LENK
STEPHEN ROSS BUSINESS SCHOOL AT THE UNIVERSITY OF MICHIGAN
KATYA SERYAKOVA
KNOWLEDGE NETWORKS, INC.
ELLEN VECCIA
KNOWLEDGE NETWORKS, INC.

ABSTRACT

A major limitation of MaxDiff scaling or any discrete-choice conjoint methods such as choice-based conjoint (CBC) is the loss of a common origin across subjects. In these models, subjects' preferences are measured relative to a base option, which eliminates a common origin for making between-subjects comparisons. This assumption allows the ranking of options within a subject, but invalidates sorting subjects by the intensity of their preferences. We propose augmenting discrete-choice data with ratings data in order to recover the common origin. We fuse the two sources of information with a joint model that contains common parameters for the discrete-choice task and ratings scales. In particular, the partworths in the MaxDiff task enter the model for the ratings data, and the identification constraints are placed on the ratings model instead of the MaxDiff model. We demonstrate that the proposed method extends the range of applications for MaxDiff and CBC.

Acknowledgements: We thank John Wurst for his helpful and insightful comments on this article and our 2007 Sawtooth Software Conference presentation. We also acknowledge the support of Sawtooth Software.

1. INTRODUCTION

MaxDiff (AKA best/worst analysis) is a special kind of scaling methodology that has been steadily gaining popularity in recent years. Initially described by Finn & Louviere (1992), it exploits the ability of subjects to pick out extreme cases – the most and least preferred options – from sets of alternatives. The sets are based on an experimental plan allowing functions of alternatives' partworths to be estimated in the same manner as partworths are estimated using traditional or choice-based conjoint data. The procedure balances the effort required by the subject in the elicitation task with the amount of information provided by the task. The most informative, non-metric measurement method requires subjects to fully rank the options in each choice set. However, for more than around five options, full rankings are notoriously unreliable and taxing. At the other extreme, pick-best is the easiest for subjects, but the least informative for estimation. Best/worst effectively doubles the sample information over pick-best, while not requiring much additional effort from subjects. The majority of MaxDiff applications to date have been with atomic options, such as selecting the most and least important attribute; however, MaxDiff can also be applied to composite options, such as product descriptions that consist of multiple attributes.

Since its introduction, this technique has been extended or enhanced in various ways, making it even more useful. The application of MCMC methods for estimating the parameters of Hierarchical Bayes models provides more accurate and informative score estimates at the individual level than the analysis method originally proposed. The development of algorithmic procedures for designing experiments has made it possible to employ experimental plans for MaxDiff choice sets that are good compromises between the number of choice sets and the precision of partworth estimates. Recent theoretical work has examined whether MaxDiff data are consistent with Random Utility Theory (Marley & Louviere, 2005). MaxDiff tasks have been shown to be a desirable means of generating results useful for market segmentation (Cohen, 2003; Cohen & Niera, 2003) as compared to other ways of collecting survey data.

A limitation of MaxDiff scaling, and choice-based conjoint (CBC) methods in general, is subjects' partworths are not measured on a common scale. Each survey-taker's scale scores are relative to an arbitrary origin, but that origin may or may not be the same for different survey-takers on the scale that is presumed to underlie their choices. The task does not capture sufficient information for estimating scores on a scale with the same origin for all survey-takers. Operationally, the preference for a base option or one of the intercepts for each subject is arbitrarily assumed to be zero in order to identify the model. This constraint shifts each subject's scale to zero at the base option, and the other partworths are measured relative to the base option. So, for example, if the scale is derived from "Most" and "Least" importance selection, one person's scale value of 2.5 for Option A means that this subject rates the option 2.5 above the base option. Another person's value of 2.5 for Option A only reflects the same relative distance from the base option, and does not imply that the two subjects would view Option A with the same absolute importance.

The loss of a common origin does not impede the application of MaxDiff to marketing applications where only within-subjects preferences are required. A common application is to use the estimated partworths from MaxDiff or discrete-choice as input to market share simulators. Since market share simulators are only concerned with the relative ranking of options within each subject, the loss of inter-subject comparability does not impact the derived market shares. However, the loss of a common origin does affect MaxDiff's usefulness in applications that require inter-subject comparisons, such as segmentation and targeting. Because the partworths are not measured on a common scale, it is not possible to sort subjects based on their preferences. Two subjects may give the same preference for Option A relative to the base option, but it is impossible to infer which subject prefers Options A the most.

Böckenholt (2004) considered this issue in the context of paired comparison choices, and proposed three methods of recovering a common origin. The researcher could a priori specify the common origins. Böckenholt gave the example of choosing among gambles with monetary winnings where the base option is the subject's current wealth. Using the subjects' actual current wealth, one is able to compare utilities for the gambles across subjects. This approach has limited application because it can only be used when the researcher is willing to make very strong assumptions about the origin.

His second approach is very creative and novel: it uses comparisons among bundles of options to recover the common origin. A hypothetical task would be to choose between an 80 GB video iPod™ or a 30 GB Zune™ digital media player bundled with a MP3 portable stereo player. A major, critical assumption of this approach is that the utilities are additive: the utility of the bundle is the sum of the utilities of its components. To see how bundles can resolve the origin, assume that U_1 , U_2, and U_3 are the latent utilities for the iPod™; the Zune™, and the portable stereo player, respectively, and that the subject evaluates three choice tasks:

1. Choose between the iPod™ and the Zune™.
2. Choose between the iPod™ and the stereo player
3. Choose between the iPod™ and the bundle.

Task 1 provides information about $U_1 - U_2$; Task 2 provides information about $U_1 - U_3$; and Task 3 provides information about $U_1 - U_2 - U_3$, assuming additivity. Then the contrast Task 1 + Task 2 – Task 3 provides a pure estimate of U_1. Without Task 3, one could only estimate the relative utilities $U_1 - U_2$ and $U_1 - U_3$.

This approach can be implemented with Sawtooth Software's CBC/HB package if one is able to make an additional assumption about the error terms of the random utilities. The appropriate random utility model (RUM) would be the following:

$Y_1 = U_1 + \varepsilon_1$ is the random utility for the iPod™.

$Y_2 = U_2 + \varepsilon_2$ is the random utility for the Zune™.

$Y_3 = U_3 + \varepsilon_3$ is the random utility for the portable stereo player.

$Y_4 = U_2 + U_3 + \varepsilon_4$ is the random utility for the bundle.

Then the important assumption behind CBC/HB is that the error terms $\{\varepsilon_j\}$ for options in a choice set are a random sample from an extreme value distribution, which leads to the standard, logistic probabilities. A more natural assumption is to write the utility of the bundle as

$Y_4 = U_2 + U_3 + \varepsilon_2 + \varepsilon_3$,

which assumes that both utilities are additive in both their deterministic and random components. Conceptually, it may be a stretch to assume that the deterministic components U_2 and U_3 are additive in (4) but not the random components, as in (5). Nevertheless, the above specification (1)-(4) may provide a fruitful and practical method for identifying the origin with standard software. It's worth noting that, aside from the strong assumption of additivity required, applications of this bundling approach are limited to those in which the alternatives being scaled can actually be combined, i.e. that aren't mutually exclusive for some reason. Preferences for a single vacation destination, type of first job, case color for an MP3 player, or flavor of ice cream are examples of alternatives that can't be bundled in a way that is likely to make sense to research subjects or users.

Böckenholt's third proposal is to augment the MaxDiff or discrete-choice data with information collected on a continuous scale. For instance, combining importance ratings on a 5 point Likert scale with the MaxDiff task. In general, we believe that this approach is superior to comparing bundles of options because it avoids the additive utility assumption for the bundle. If the bundles are not additive, then the procedure produces systematic bias in the estimates of the absolute utilities. On the other hand, fusing MaxDiff with ratings has its own challenges. First, to use ratings and discrete-choice scales to identify the origin, the model for the ratings must include the partworths used in the MaxDiff or CBC task. Second, the model for ratings data needs to accommodate well-known scale usage bias (Rossi, Gilula, and Allenby 2001) or else the imputed origin may only reflect scale usage. Finally, a large body of literature in psychometrics documents failure of procedure invariance in the elicitation of preferences (c.f. Slovic 1995). For example, Lichtenstein and Slovic (1971) demonstrated preference reversals for pairs of gambles depending when subjects are asked to price the gamble than choose the most preferred gamble. Grether and Plott (1979) systematically investigated a number of potential "rational" explanations for preference reversals and concluded that preference reversals arise from different psychological processes for the different tasks, despite their concerted attempts to prove the contrary. Consequently, any model that attempts to fuse ratings and discrete choice needs to be sufficiently flexible that the psychological processes do not distort the partworths in the discrete choice task. Here, we are making the value judgment that the discrete-choice task provides better external validity than ratings because customers in the market place choose products and do not rate them.

It is worth noting that a method that does fuse preference or importance responses obtained using different elicitation methods may provide more stable, and more generalizable results compared to any single elicitation technique as it is in a sense integrating over tasks that may each create its own method-specific bias. Also, when you use just a single task, you don't have the opportunity to observe failure of procedural invariance. That doesn't mean that it wouldn't occur, of course.

The rest of the paper presents the model for fusing ratings and discrete-choice data in order to recover a common origin and demonstrates its utility in targeting subjects with two examples. In the next section, we review the underlying random utility models (RUM) for discrete-choice and MaxDiff. These models assume that the observed choices are driven by unobserved utilities for the options in a choice task. We then extend the basic RUM to fuse ratings and discrete-choice data in both the logit and probit specifications in Section 3. Section 4 reports the findings from a simulation study and demonstrates using the common origin for targeting subjects, and Section 5 concludes the paper.

2. Random Utility Models for Discrete Choice and MaxDiff

2.1 Random Utility Essentials and the Loss of the Common Origin

Since McFadden's (1974) seminal work on economic choice, random utility models (RUM) have provided the foundation for discrete-choice experiments. The concept is very simple: subject i has a latent, random utility Y_{ik} for option k. This utility is called "latent" because the researcher does not observe it directly. Instead, he or she can only observe its consequences, namely choices, best/worst, or rankings. The random utility is decomposed into a deterministic component U_{ik} and random component ε_{ik}: $Y_{ik}=U_{ik} + \varepsilon_{ik}$. If the random components have an extreme-value or Gumbel distribution[1] McFadden (1974) showed that the choice probabilities are logistic functions of the partworths, which is the underlying assumption of Sawtooth Software's CBC/HB software. If the random components are normally distributed, then the choice probabilities are probit functions (Aitchison and Bennet 1970). The type of task determines the likelihood function that links the observed data U_{ik} by integrating over the random component. This likelihood function varies for pick-best, full rankings, and MaxDiff due to different processes for using the random utilities to generate the response.

If the task is pick-best, then each subject is presented with J different choice tasks where each choice task has K options. Now we have three subscripts, i for subject, j for choice task, and k for option within choice task, and $Y_{ijk} = U_{ijk} + \varepsilon_{ijk}$ is the corresponding random utility. The behavioral assumption that makes pick-best work is that the subject chooses that option that corresponds to the maximum latent utility:

Pick option s if $Y_{ijs} \geq Y_{ijk}$ for all options k in the choice set. (1)

The choice probabilities as functions of U_{ijk} are then computed using this inequality by integrating over the random terms.

The condition in (1) relates the observed choice to the preference structure, and focuses our attention on the loss of the common origin. The inequality in (1) holds for any linear rescaling of the latent utilities: nothing changes in the preference structure if subject i uses $Y^*_{ijk} = aY_{ijk} + b$ for arbitrary constants a and b where a is a positive. The researcher (or software package) uniquely identifies the latent utility by imposing constraints on the latent utilities. Most commonly, the scale parameter for the extreme value distribution (in logit models) or one of the variances for normal distribution (in probit models) is fixed to one. This constraint forces a to be one in $Y^*_{ijk.}$ To eliminate arbitrary scale origins, one of the U_{ijk} is set to 0, say $U_{ij1} = 0$, which is the same as measuring the utilities of the other options with respect to option 1. If the options are composite options, $U_{ijk} = x_{ijk}'\beta_i$, where x_{ijk} is a vector describing attribute levels and β_i is a vector of partworths, then one of the intercepts is set to 0. Therein lies the crux of the problem: to estimate preferences uniquely from discrete choice data, the researcher loses the common scaling and inter-subject comparability.

To complete our review of random utility models, if the random terms are a random sample from right-skewed, extreme value (a.k.a. Gumbel) distributions with scale parameter 1, the choice probabilities are a logistic function of the $\{U_{ijk}\}$:

$$P_{ij}\left(s = \text{Most Preferred}\right) \equiv P_{ij}\left(s\right) = \frac{\exp\left(U_{ijs}\right)}{\sum_{u=1}^{K}\exp\left(U_{ijk}\right)} \text{ for } s = 1,\ldots,K\,. \quad (2)$$

Again, it is evident that one could add an arbitrary constant to each U_{ijk} without altering the choice probability unless an identifying constraint in enforced.

[1] The cumulative distribution function is $F(\varepsilon) = \exp[-\exp(-\varepsilon)]$ for the right-skewed, extreme value distribution.

2.2 Random Utility Models for MaxDiff

MaxDiff, originally proposed by Louviere and Woodworth (1990) and published by Finn and Louviere (1992) applies the basic RUM to best/worst responses. Their model assumes that subjects evaluate the difference in utility for every pair of options and selects the difference with maximal utility. The random utility for each ordered pair (s,t) in choice task j is:

$$Y_{ij,st} = U_{ijs} - U_{ijt} + \varepsilon_{ij,st} \quad \text{for } s, t = 1, \ldots, K \text{ and } s \neq t.$$

Note that $Y_{ij,st}$ is not equal to $Y_{ij,ts}$. This MaxDiff model assumes that the option with the maximal differences is selected:

Option s is best and option t is worst if $Y_{ij,st} > Y_{ij,uv}$ for all other pairs (u,v).

Assuming extreme value distributions for the random terms, the MaxDiff probabilities for subject i and choice task j are:

$$P_{ij}\left(s = \text{Most Preferred}, t = \text{Least Preferred}\right) \equiv P_{ij}\left(s,t\right)$$
$$= \frac{\exp\left(U_{ijs} - U_{ijt}\right)}{\sum_{u=1}^{K} \sum_{v=1:v\neq u}^{K} \exp\left(U_{iju} - U_{ijv}\right)} \text{ for } s \neq t \qquad (3)$$

Finn and Louviere (1992) and Flynn, Louviere, Peters, and Coast (2007) apply MaxDiff to atomic options for aggregate attitudes for food safety and quality of life. Of course, MaxDiff could also be used with composite options, in which case the choice probabilities are:

$$P_{ij}\left(s = \text{Most Preferred}, t = \text{Least Preferred}\right) \equiv P_{ij}\left(s,t\right)$$
$$= \frac{\exp\left(\left[x_{ijs} - x_{ijt}\right]' \beta_i\right)}{\sum_{u=1}^{K} \sum_{v=1:v\neq u}^{K} \exp\left(\left[x_{iju} - x_{ijv}\right]' \beta_i\right)} \text{ for } s \neq t \qquad (3')$$

Once again, the MaxDiff model is identified by assuming that one of the utilities (or intercepts for composite products) is zero.

Although the choice probabilities in (3) and (3') seem complex, this formation for MaxDiff can easily be estimated in Sawtooth Software's CBC/HB package if the number of options K in a choice task are not too large. For example, suppose that a brand study is performed with 5 brands and nominal price. To identify the model, the partworth for brand 5 is assumed to be zero. Each choice task consists of three options (K=3). In Figure 1, the choice set uses brands B1, B3, and B4, with prices \$5, \$4, and \$7, respectively. The left-side of Figure 1 gives the design matrix for this choice task where the first 4 columns identify the brand (Brand 5 is the base brand), and the last column is price. The three rows represent the three options in the choice set. The right-side of Figure 1 gives the corresponding .cho file to implement MaxDiff in CBC/HB. At the top "6 1" means that there are 6 possible choices, corresponding to the different ordered pairs of B1, B2, and B3, and 1 choice is made. Taking all possible pairwise differences of the rows of the matrix on the left-hand-side, where we added "B1-B3" etc. to indicate which brands are in the pairwise differences, forms the cho matrix on the right-hand-side. In this example, we assumed that the subject picked B3 as the best and B1 as the worst, which corresponds to the difference "B3-B1" on the right-hand-side. Consequently, the "3 99" at the bottom indicates that row 3 was selected, and 99 is a stop-code.

Figure 1:
An example of a Sawtooth Software cho matrix for the original formulation of MaxDiff where B3 is the best option and B1 is the worst option.

$$
\begin{array}{c}
\begin{matrix} \mathrm{B1}\rightarrow \\ \mathrm{B3}\rightarrow \\ \mathrm{B4}\rightarrow \end{matrix}
\begin{bmatrix} 1 & 0 & 0 & 0 & 5 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 7 \end{bmatrix}
\Rightarrow
\begin{array}{l}
6\ 1 \\
\begin{bmatrix} 1 & 0 & -1 & 0 & 5-4 \\ 1 & 0 & 0 & -1 & 5-7 \\ -1 & 0 & 1 & 0 & 4-5 \\ 0 & 0 & 1 & -1 & 4-7 \\ -1 & 0 & 0 & 1 & 7-5 \\ 0 & 0 & -1 & 1 & 7-4 \end{bmatrix}
\begin{matrix} \leftarrow \mathrm{B1\text{-}B3} \\ \leftarrow \mathrm{B1\text{-}B4} \\ \leftarrow \mathrm{B3\text{-}B1} \\ \leftarrow \mathrm{B3\text{-}B4} \\ \leftarrow \mathrm{B4\text{-}B1} \\ \leftarrow \mathrm{B4\text{-}B3} \end{matrix} \\
3\ 99
\end{array}
\end{array}
$$

A mathematically equivalent expression for the MaxDiff choice probabilities in Equations (3) or (3') is:

$$P_{ij}(s,t) \propto P_{ij}(s)Q_{ij}(t) \text{ for s, t = 1, …, K and s} \neq \text{ t.} \quad (4)$$

where $P_{ij}(s)$ is the probability of selecting the best from Equation (2) and

$$Q_{ij}(t) = \frac{\exp\left(-x'_{ijt}\beta_i\right)}{\sum_{k=1}^{K}\exp\left(-x'_{ijk}\beta_i\right)} \text{ for t = 1, …, K.} \quad (5)$$

Note, in particular, the proportionality sign in Equation (4).

A Sawtooth Software (2005) technical report derives the probabilities $Q_{ij}(t)$ in Equation (5) from RUM for the least preferred option where

$$Y^{*}_{ijk} = x'_{ijk}\beta_i + \varepsilon^{*}_{ijk} \text{ for } k = 1,\ldots,K$$

and the random terms $\{\varepsilon^{*}_{ijk}\}$ are a random sample from a *left-skewed*, extreme value distribution with scale parameter one[2]. Then option t is least preferred if $Y^{*}_{ijt} < Y^{*}_{ijk}$ for all k $\neq$ t, and the choice probability is given in Equation (4). Cohen (2003) and Cohen and Orme (2004) use a simple method for approximating the MaxDiff probabilities in Equation (3) by relaxing the constraint that the most and least preferred options have to be different:

[2] The cumulative distribution function is $F(\varepsilon) = \exp[-\exp(\varepsilon)]$ for the left-skewed, extreme value distribution.

$$P_{ij}(s,t) = P_{ij}(s)Q_{ij}(t) \text{ for s, t = 1, …, K.} \quad (5')$$

The proportionality sign in Equation (4) has been replaced by an equality sign in Equation (5'). In Equation (5'), the best and worst options are mutually independent, while they are dependent in Equation (4). Behaviorally, this formulation is different from the original MaxDiff, which assumed that the subject selected the maximum pairwise difference. This specification, which was also used by Finn & Louviere (1992) for aggregate data, corresponds to a two-stage behavioral model where the subject evaluates all of the options using right-skewed extreme value distributions and selects the best. Then he or she reevaluates all of the options with left-skewed extreme value distributions and selects the worst.

The Sawtooth Software (2005) technical report shows how to estimate the model used by Cohen and Orme using Sawtooth Software's CBC/HB and by treating the single best/worst responses as two responses in standard CBC. The best response is coded as usual with design matrix X_{ij} for subject i and choice set j. The worst response is coded with design matrix $-X_{ij}$. Figure 2 gives the cho matrix for the best/worst data in Figure 1. The top "3 1" indicates that there are 3 options in the choice set, and one option was selected. The first matrix is the standard design matrix for choice-based conjoint. The "2 99" indicates that the second row (Brand 3) was selected as "best," and "99" is a stop-code. The second "3 1" indicates, as before, three options in the choice task with one choice only. The second matrix is the coding for the "worst" choice task, which is merely the negative of the top matrix. The "1 99" indicates that the first option was selected as worst, and "99" is the stop code.

Figure 2:
The Sawtooth Software cho matrix for the Cohen/Orme best/worst model specification using the data in Figure 1.

3 1

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 5 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 7 \end{bmatrix}$$

2 99

3 1

$$\begin{bmatrix} -1 & 0 & 0 & 0 & -5 \\ 0 & 0 & -1 & 0 & -4 \\ 0 & 0 & 0 & -1 & -7 \end{bmatrix}$$

1 99

We propose an alternative formulation for best/worst data that follows directly from the basic RUM in Section 2.1. If s is the best alternative and t is the worst alternative, then

$$Y_{ijs} > Y_{ijk} > Y_{ijt} \text{ for } k \neq s \text{ or } t, \text{ and } k = 1, …, K. \quad (6)$$

Here, we assume that the random terms are either a random sample from a right-skewed distribution (standard logit model) or a normal distribution (probit model). The behavioral assumption is that the subject evaluates the latent utility for each option and selects the options with the maximum and minimum utilities. The reader may wonder why other authors have not thought of (6). The sad news is that the specification in (6) does not lead to a tidy expression for the choice probabilities. However, a close inspection of Equation (6) reveals that best/worst responses correspond to partially ranked data. Our approach to the problem is to impute the missing ranks and use the exploding logit model of Beggs, Cardell and Hausman (1981) and Chapman and Staelin (1982).

In particular, for subject i and choice task j, let $R_{ij1, \ldots,} R_{ijK}$ be the ranks of the options where R_{ij1} is the index of the least preferred option and R_{ijK} is the index of the most preferred options. If all of the options in the choice task were ranked ordered by the subject, then $Y_{ijR1} < Y_{ijR2} < \ldots < Y_{ijRK.}$ The exploding logit model for fully ranked data is:

$$P\left(R_{ij1},\ldots,R_{ijK}\right)=\prod_{s=2}^{K}\frac{\exp\left(U_{ijR_s}\right)}{\sum_{k=1}^{s}\exp\left(U_{ijR_K}\right)} \text{ for } R_{ij1}<\cdots<R_{ijK} \tag{7}$$

The marginal distribution for the best and worst options can be obtained from Equation (7) by summing over the intervening ranks. Since these models are estimated by using MCMC, an alternative to direct computation is to impute the missing ranks in the MCMC algorithm by the following procedure. Given all of the parameter estimates, generate the latent utilities Y_{ijk} for k not equal to R_{ij1} or R_{ijK}:

$Y_{ijk} = U_{ijk} - \ln[-\ln(u)]$ where u is uniform on [0,1] for $k \neq R_{ij1}$ or R_{ijK}.

Then the missing ranks are based on these imputed latent variables. For the probit formulation, one merely adds an additional constraint: the lower inequality in Equation (6), to the standard MCMC algorithm of Albert and Chib (1993) and McCulloch and Rossi (1994).

2.3 Simulation Study Comparing MaxDiff Methods

At this point, a natural question is if these three procedures have any practical differences. Their underlying behavioral models start with RUM. However, the processes of using the random utilities to select the best/worst alternatives are qualitatively different, and the resulting likelihood functions are different. We use a short simulation study to see if there are quantitative differences in the results. The study simulates 100 subjects who evaluate 26 to 36 choice tasks. Each choice task has 4, full profile options. Each profile corresponds to a brand (4 brands). The profiles also include price X1 (continuous scale) and a 0/1 dummy X2 for advertising. The individual partworths $\{\beta_i\}$ are related to subject-level demographics – ln(income) and household size – through a multivariate regression model. Table 1 compares the true values of $\{\beta_i\}$ to their estimates using the original MaxDiff (Equations 3 and 3'), the Finn & Louviere / Cohen & Orme specification "LR Skew" (Equation 5), and the rank imputed exploding logit "RIMEX" (Equations 6 and 7). Based on this simulation's result there is nothing that distinguishes one approach over the other. Figure

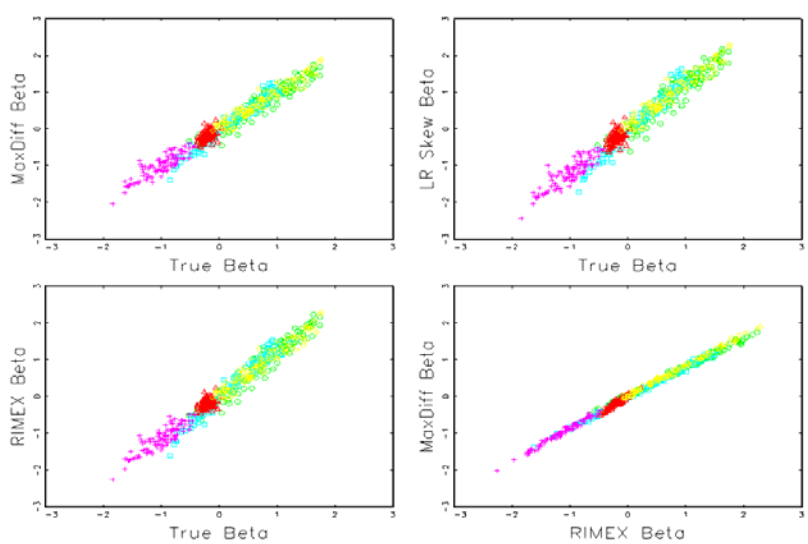
3 provides a graphical display of Table 1 by plotting the true and estimated partworths for the three procedures.

Table 1.
Simulation results for comparing the estimated partworths of the original MaxDiff, procedure of Cohen/ Orme ("LR Skew"), and the rank imputed exploding logit ("RIMEX").

Correlation	Brand 1	Brand 2	Brand 3	X1	X2
True vs RIMEX	0.965	0.979	0.386	0.876	0.967
True vs MaxDiff	0.955	0.978	0.433	0.866	0.966
True vs LR Skew	0.952	0.978	0.409	0.865	0.967
RIMEX vs MaxDiff	0.997	0.997	0.937	0.989	0.997
RIMEX vs LR Skew	0.997	0.997	0.944	0.991	0.997
MaxDiff vs LR Skew	0.999	1.000	0.994	0.997	1.000
RMSE	Brand 1	Brand 2	Brand 3	X1	X2
True vs RIMEX	0.284	0.282	0.119	0.224	0.219
True vs MaxDiff	0.182	0.172	0.125	0.161	0.129
True vs LR Skew	0.294	0.323	0.153	0.272	0.249
RIMEX vs MaxDiff	0.262	0.172	0.049	0.164	0.172
RIMEX vs LR Skew	0.065	0.089	0.075	0.073	0.062
MaxDiff vs LR Skew	0.244	0.189	0.066	0.210	0.204

RMSE is root mean squared error.

Figure 3.
True and estimated partworths from the simulation study.



We also ran other simulations that varied the number of subjects and number of choice tasks per subject and obtained similar conclusions that the estimates from the three approaches did not systematically differ, despite our concerted effort to show the contrary. Our results are not bad news for practitioners, as they suggest that current MaxDiff modeling practices provide reasonably good results, ceteris paribus.

3. FUSING MAXDIFF WITH RATINGS TO IDENTIFY THE ORIGIN

This section augments the best/worst or discrete-choice data with ratings data to identify the origin for measuring partworths. In describing a model to fuse discrete-choice and ratings data, we assume that the partworths from the discrete-choice model are the focal set of parameters. Moreover, we do not want the psychological processes used for ratings to systematically distort the preference structure from the choice task. We assume that the discrete-choice task has more external validity than the ratings task. Our main purpose of fusing the choices and ratings is to identify the common scale for inter-subject comparisons. Other objectives, such as better estimates, are secondary concerns and not investigated here. However, we need a model that uses the ratings data to recover the origin without contaminating the discrete-choice partworths. A more detailed description of the model is given in Lenk and Bacon (2007), where it is noted that the data fused with choice data need not be ratings data. They could be behavioral measures based on purchase histories or web site visits, for example. Our model specification is related to that in Vriens, Oppewal, and Wedel (1998) who fused a ratings conjoint task with importance ratings to obtain better partworth estimates.

We use a threshold model (Aitchison and Silvery 1957 and Rossi, Gilula, and Allenby 2001) to relate the ratings data to a latent variable, and then we specify a joint model for the latent variables used in the ratings and discrete-choice tasks. The threshold model converts the ordinal responses on a rating scale to a continuous scale. It assumes that the observed ordinal responses arise due to the continuous latent value falling in regions determined by ordered cutpoints C_1 to C_H for an H point scale. The model for the latent variables is:

$$W_{im} = \varphi_i + \alpha_m + v'_{im}\Psi\beta_i + \xi_{im} \qquad (8)$$

where

1. φ_i is a random effect for subject i. The random effects are random sample from a normal distribution with mean 0 and standard deviation τ. These parameters help ameliorate scale usage effects.

2. α_m is a parameter for item m that adjusts the ratings model for compatibility effects.[3]

3. v_{im} is the design vector for item m and is observed. For atomic options, e.g. importance ratings, it is a vector of zeros and a single one.

4. Ψ is a square, diagonal matrix with zeros on the off-diagonals and positive entries on the diagonals. Ψ adjusts the partworths for prominence effects[4] when going from discrete-choice to ratings tasks. Its elements allow making inferences about the appropriateness of fusing the choice and ratings data.

5. β_i are the partworths from the discrete-choice or MaxDiff task.

6. ξ_{im} is a random term, which is either right-skewed extreme value for the logit model or normally distributed for the probit model. The scale parameters (logit) or error variances (probit) depend on the item m.

---

[3] Compatibility effects occur when a stimulus is more compatible with a particular elicitation task. See Tversky, Sattath, and Slovic (1988).

[4] Different attributes become more or less prominent depending on the task. See Nowlis and Simonson (1997).

The probabilities for the item responses are:

$$P_{im}(1) = F\left[\frac{C_1 - (\varphi_i + \alpha_m + v'_{im}\Psi\beta_i)}{\zeta_m}\right]$$

$$P_{im}(h) = F\left[\frac{C_h - (\varphi_i + \alpha_m + v'_{im}\Psi\beta_i)}{\zeta_m}\right] - F\left[\frac{C_{h-1} - (\varphi_i + \alpha_m + v'_{im}\Psi\beta_i)}{\zeta_m}\right].$$

$$for\ h = 2, \ldots, H-1$$

$$P_{im}(H) = 1 - F\left[\frac{C_H - (\varphi_i + \alpha_m + v'_{im}\Psi\beta_i)}{\zeta_m}\right]$$

where $C_1 < \ldots < C_H$ are the cutpoints. In the logit model, F is the cumulative distribution function for the right-skewed, extreme value distribution in footnote 4, and ζ_m is the scale parameter for item m. In the probit model, F is the cumulative distribution function of the standard normal distribution, and ζ_m is the error standard deviation for item m. This model for the ratings includes the partworths from the choice task in an indirect fashion to accommodate potential psychological biases when going from discrete-choice tasks to rating tasks. When these biases are absent, then α_m is zero and the Ψ is the identity matrix.

The RUM for the discrete-choice or Max/Diff task is:

$$Y_{ijk} = x'_{ijk}\beta_i + \varepsilon_{ijk}$$

$$\beta_i = \Theta' z_i + \delta_i$$

where

1. Y_{ijk} is subject i's latent utility for profile k in choice task j
2. x_{ijk} is the design vector for profile k.
3. ε_{ijk} is the random term either from a right-skewed, extreme value distribution (logit model) or multivariate normal distribution (probit model).
4. z_i is a subject-level covariate; Θ is a matrix of regression coefficients, and δ_i is a multivariate normal error term with mean 0 and covariance Λ.

We use standard priors for the parameters (See Lenk and Bacon 2007). For pick best data, we assume the constraint (1). For best/worst data, we impose the constraint in Equation (6). The model could be easily modified for the original MaxDiff formulation or the Cohen/Orme approach.

The key to recovering the common origin is moving the standard identification constraints from the choice task to the ratings model. We do not constrain the partworths $\{\beta_i\}$, and we do not assume that the scale parameter (logit) or a variance (probit) for the random term $\{\varepsilon_{ijk}\}$ is one. Instead, we assume that one of the $\{\alpha_m\}$ is zero, and one of the diagonal elements of Ψ is one.

4. Applications

Lenk and Bacon (2007) present simulation results that indicate the model in Section 4 is identified, and the estimated parameters are close to their true values. Here, we present two applications that illustrate the practical benefits retaining a common origin for subjects' utility scales in order to compare subjects on their preference structures.

4.1 Educational Goals

A study on the importance of eight educational goals had 1470 subjects both rate the importance for each goal, and then perform a best/worst task consisting of 8 choice sets with three goals per choice set. Without ratings data, we identify the model by assuming that the utility for Goal 8 is zero. We fitted the model with just the best/worst data and with both the best/worst and ratings data. Also, we fitted both logit and probit models, which gave similar results. We will report the results for the probit model. We will not report all of the estimates of the parameter in Section 4, and we will only focus on the estimated partworths.

Figure 4 plots the estimated partworths for the two models: the x-axis is for the standard MaxDiff model without ratings and the y-axis is our proposed model. The graph shows that with a common origin, subjects are distinguished by the absolute utility evaluations. We emphasize this point by segmenting the subjects into three tiers. The top tier consists of subjects with three or more partworths that are greater than 6, and the bottom tier consists of subjects with three or more partworths less than 1.5. There are approximately 20% of the subjects in each of the top and bottom tiers. The top tier represents subjects who are highly concerned with education, while the bottom tier is not. Figure 5 presents boxplots of the partworths for the three tiers with ratings and without ratings. The boxplots show the partworths increasing through the tiers when they are estimates with the ratings, but they are flat when estimated without ratings.

Figure 4.
Estimated partworths for educational goals with and without ratings.

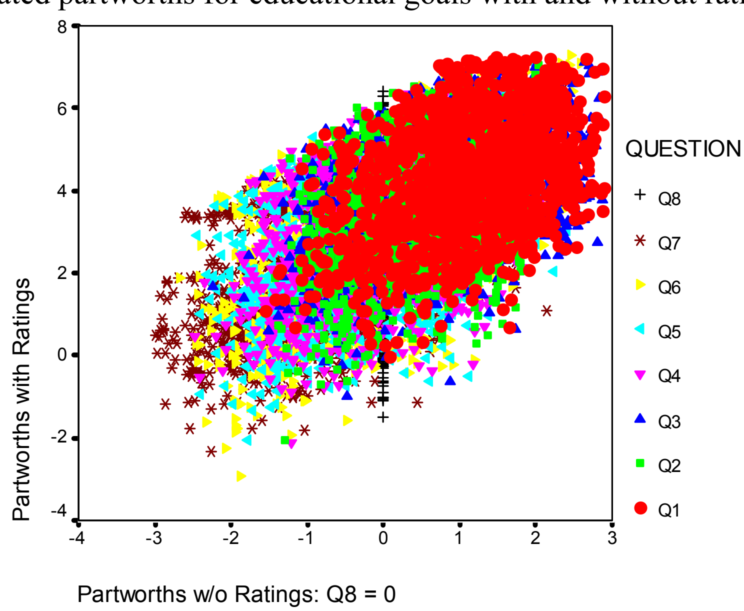
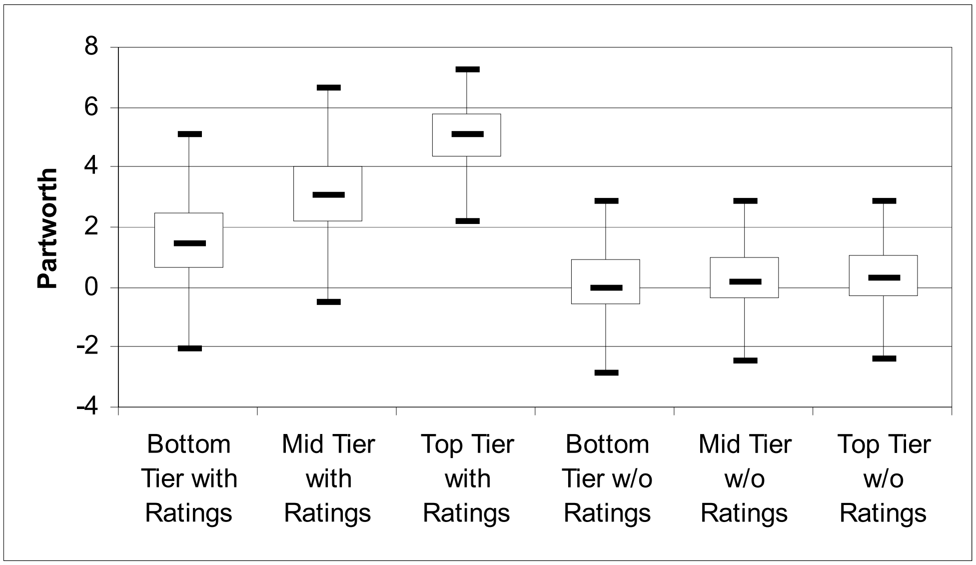


Figure 5
Boxplots of partworths segmented by tiers with and without ratings.



4.2 PC Design

The second application uses data from a conjoint study of personal computers. The participants were 210 MBA students at a major university. These subjects rated 20 profiles on a 0 to 10 Likert scale for the likelihood of purchase, and for each profile, they also indicated Buy/No Buy. The data are taken from Lenk *et al.* (1996). We use a threshold model for the discrete choice task where subject i selects "Buy" for profile j if his or her random utility $Y_{ij} > \beta_{iT.}$ The threshold can be interpreted as the utility of the "outside good." For example, if the subject already owns a PC, then β_{iT} can be interpreted as the utility for that PC. In standard binary choice models without ratings, the utility of the outside good is assumed to be 0. In our fusion model, we are able to separately estimate both the utility of the outside good and the intercept, which is the utility of the "average" PC in the study since we use effects coding for the binary attributes.

Figure 6 displays boxplots of the partworth heterogeneity for the models with and without the importance rating data. With importance rating (right-hand panel), the partworth for the outside good (the threshold) and the "average PC" (the intercept) are separately estimated, while they are confounded for the model without ratings. Figure 7 sorts subjects according to their utility of the outside good and superimposed the price partworth. The correlation between the partworths for price and the outside good is –0.376. The plot shows that the subjects with smaller utilities of outside goods also tend to be less price sensitive, and conversely. This implies that subjects who own a fairly good PC require an outstanding deal to buy a new PC, while subjects with low utility for the outside good are willing to pay more.

Figure 6.
PC design study: boxplots of the heterogeneity in the estimated partworths .
The estimates without ratings is the left-hand panel, and the model with ratings is the right-hand panel.

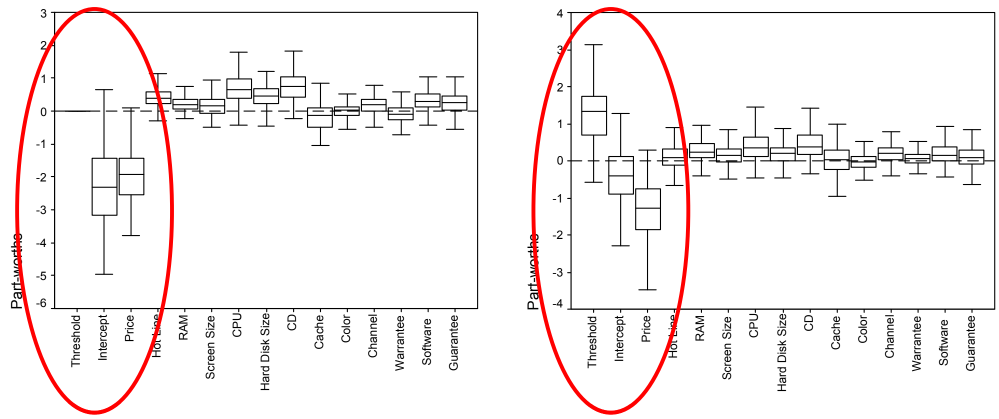
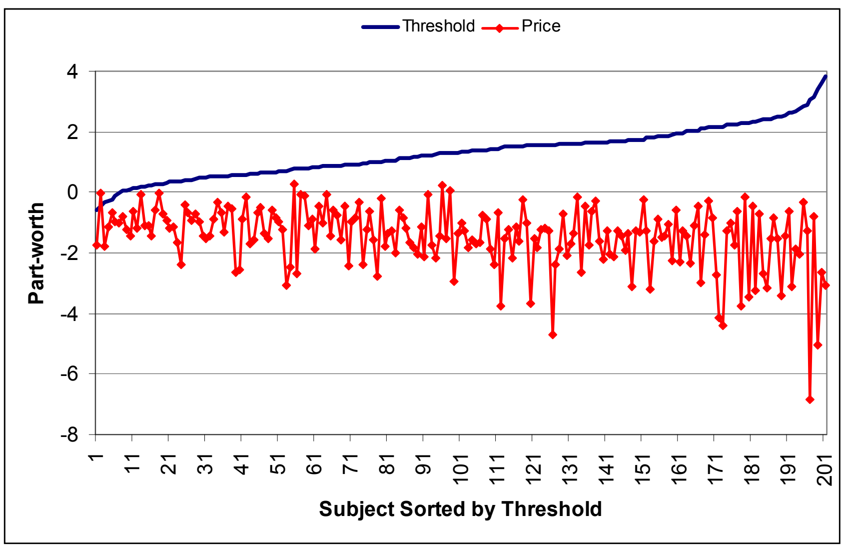


Figure 7.
PC design study: subjects are sorted by their utilities for the outside good.



5. Conclusion

The purpose of this paper is twofold: to make marketing researchers aware of an overlooked limitation of choice-based models, the loss of a common origin, and to provide a remedy by augmenting choice data with ratings data. Without a common origin inter-subject comparisons are not meaningful. Applications such as market share simulations, which only rely on the relative partworths within each subject, are still valid without a common origin. However, applications such as segmentation and targeting that rely on comparisons of preference structures among subjects are misleading unless partworths are measured on a common scale. We propose augmenting choice-based data with auxiliary data, which in the present case are ratings data, to recover a common scale. Our fusion model accommodates psychological biases inherent in ratings, as compared to choices, and scale usage biases for ratings. In two datasets we demonstrate that recovering a common origin increases the utility of choice-based experiments.

REFERENCES

Albert J H and Chib S (1993). "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88 (422), Jun, 669-679.

Aitchison J and Bennet JA (1970). "Polychotomous quantal response by maximum indicant," *Biometrika*, 57, 253-262.

Aitchison J and Silvery SD (1957). "The generalization of probit analysis to the case of multiple responses," *Biometrika*, 44, June, 131-150.

Beggs S, Cardell S and Hausman J (1981). "Assessing the potential demand for electric cars," *Journal of Econometrics*, 17, l-20.

Böchenholt, U. (2004). "Comparative judgments as an alternative to ratings:Identifying the scale origin," *Psychological Methods*, 9 (4), 453-465.

Chapman, R G and Staelin R (1982). "Exploiting rank ordered choice set data within the stochastic utility model," *Journal of Marketing Research*, 19 (3), Aug, 288-301.

Cohen, Steve (2003). "Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation," *2003 Sawtooth Software Conference Proceedings*, Sequim, WA.

Cohen, S. & Orme, B. (2004). "What's your preference?" *Marketing Research Magazine,* Summer, 32-37.

Cohen, S. & Neira, L. (2003). "Measuring preference for product benefits across ountries: Overcoming scale usage bias with maximum difference scaling." Paper presented at the Latin American conference of the European society for opinion and marketing research, Punta del Este, Uruguay (pp. 1–22). Reprinted in *Excellence in International Research*: 2004. ESOMAR, Amsterdam, Netherlands, 2004.

Finn A and Louviere J J (1992). "Determining the appropriate response to evidence of public concern: The case of food safety," *Journal of Public Policy & Marketing*, 11 (1), 12-25.

Flynn TN, Louviere J J, Peters T J, and Coast J (2007). "Best-worst scaling: What it can do for health care research and how to do it," Journal of Health Economics, 26, 171-189.

Grether, D. M. and Plott, C. (1979). "Economic theory of choice and the preference reversal phenomena," *American Economic Review*, 69, 623-638.

Lenk, P and Bacon, L (2007). "Estimating Common Utility Origins and Scales in Discrete-Choice Conjoint with Auxiliary Data," working paper, The University of Michigan.

Lenk, P., DeSarbo, W., Green, P., and Young, M. (1996). "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15 (2), 173-191.

Lichtenstein, S. and Slovic, P. (1971). “Reversal of preferences between bids and choices in gambling decisions,” *Journal of Experimental Psychology*, 89, 46-55.

Louviere J and Woodworth G (1990). “Best-Worst scaling: A model for largest differences judgements,” working paper, Faculty of Business, University of Alberta.

Marley, A. A. J. and Louvier, J. J. (2005). “Some probabilistic models of best, worst, and best-worst choices,” *Journal of Mathematical Psychology*, 49, 464-480.

McCulloch, R and Rossi P E (1994). “An exact likelihood analysis of the multinomial probit model,” *Journal of Econometrics*, 64 (1-2), Sept-Oct, 207-240.

McFadden D (1974). “Conditional logit analysis of quantitative choice behavior,” Frontiers in Econometrics, ed by P Zarembka, Academic Press, New York, 105-142.

Nowlis, S. M. and Simonson, I. (1997). “Attribute-task compatibility as a determinant of consumer preference reversals,” *Journal of Marketing Research*, 34 (2), May, 205-218.

Rossi, P. E., Gilula, Z., and Allenby, G. M. (2001). “Overcoming scale usage heterogeneity: A Bayesian hierarchical approach,” *Journal of the American Statistical Association*, 96 (453), Mar, 20-31.

Slovic, P (1995). “The Construction of Preference,” American Psychologist, 50 (5), May, 364-371.

Sawtooth Software (2005). *The MaxDiff/Web System Technical Paper,* Sawtooth Software, Inc.

Tversky, A., Sattath, S., and Slovic, P. (1988). “Contingent weighting in judgment and choice,” *Psychological Review*, 95 (3), July, 371-384.

Vriens, M., Oppewal, H., and Wedel, M. (1998). “Ratings based versus choice-based latent class conjoint models - An empirical comparison,” *Journal of the Market Research Society*, 40 (3), July, 237-248.