Web Appendices for
Unstructured Direct Elicitation of Decision Rules

Min Ding

John Hauser

Songting Dong

Daria Dzyabura

Zhilin Yang

Chenting Su

Steven Gaskin

February 2010

These appendices are provided as supplements to "Unstructured Direct Elicitation of Decision Rules." The appendices are available from the authors and on the *Journal of Marketing Research* website.

Min Ding is Smeal Research Fellow in Marketing and Associate Professor of Marketing, Smeal College of Business, Pennsylvania State University, University Park, PA 16802-3007; phone: (814) 865-0622; fax: (814) 865-3015; minding@psu.edu.

John R. Hauser is the Kirin Professor of Marketing, MIT Sloan School of Management, Massachusetts Institute of Technology, E40-179, One Amherst Street, Cambridge, MA 02142, (617) 253-2929, fax (617) 253-7597, hauser@mit.edu.

Songting Dong is a Lecturer in Marketing, Research School of Business, the Australian National University, Canberra, ACT 0200, Australia, dongst@gmail.com.

Daria Dzyabura is a doctoral student at the MIT Sloan School of Management, Massachusetts Institute of Technology, E40-170, One Amherst Street, Cambridge, MA 02142, (617) 253-2268, dariasil@mit.edu.

Zhilin Yang and Chenting Su are both Associate Professor of Marketing, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong; phone: (852) 3442-4644; fax (852) 3442-0346; mkzyang@cityu.edu.hk, mkctsu@cityu.edu.hk.

Steven P. Gaskin is a Principal at Applied Marketing Science, Inc., 303 Wyman Street, Waltham, MA 02451, (781) 250-6311, sgaskin@ams-inc.com.

WEB APPENDIX 1

ANALYSIS OF CHOICE WITHIN THE CONSIDERATION SET

Our focus in the paper is on consumers' consideration-set decisions. We chose this focus for managerial and scientific interest because it enabled us to test a range of non-compensatory and compensatory decision rules and because, for categories with many products and many features, consideration is an important managerial problem. The focus also simplified exposition.

Our studies also asked respondents to rank profiles within their consideration sets. For mobile phones, three of the four decompositional methods rank the profiles and both direct-elicitation methods weakly rank the profiles. From these predicted ranks we compute the rank correlation with the observed ranks within the consideration sets in the validation data. Table A1 summarizes the results.

Table A1 is consistent with Table 1 in the text. There is no statistical difference between the decompositional additive logit method and the unstructured direct-elicitation (UDE) method on both the initial and the delayed validation. The greedoid dynamic program does not do as well on choice as consideration, possibly because non-compensatory models are more common in consideration than choice – an hypothesis worth further testing.

TABLE A1. RANK CORRELATIONS FOR CHOICE WITHIN CONSIDERATION SET

Mobile Phone Study	***Initial Validation***	***Delayed Validation***
Decompositional Methods		
HB Logit, Additive Utility	**0.374***	**0.396***
HB Logit, *q*-Compensatory	0.346	0.328
Greedoid Dynamic Program[1]	0.268	0.273
Direct-Elicitation Methods		
Structured direct elicitation	0.332	0.267
Unstructured direct elicitation	**0.412***	**0.375***

[1] Estimates a lexicographic model. * Best or not significantly different than best at the 0.05 level.

For the automotive study, UDE is best or not significantly different than best. However, unlike for consideration decisions, UDE is not significantly better in predicting ranks within the consideration set than the decompositional methods. We are hesitant to read too much into this result because the variation across respondents in rank correlations is large compared to variation across methods. The ratio of the standard deviation to the mean is between 2.0 and 2.7 for the four methods (although the results in Table A2 are paired *t*-tests with greater power).

The lack of statistical power for ranks in the automotive study is explained, in part, because we focused that study on consideration-set decisions. In the current managerial climate, understanding automotive consideration is extremely important. The large number of potential features and levels (53) relative to the sizes of the consideration sets (~ 10 profiles) challenged all methods. Nonetheless there remains the scientific challenge of improving and testing UDE for automotive ranks within the consideration set. For example, more-aggressive coding might resolve ties in weak orderings to improve the predictive ability of UDE for ranks.

An alternative hypothesis is that UDE is best for heuristic consideration decisions while additive decomposition is best if compensatory rules are used to rank profiles within a consideration set. (See also the greedoid dynamic program results in Table A1.) We cannot resolve this hypothesis with our focused studies, but it is an interesting topic for future research.

TABLE A2. RANK CORRELATIONS FOR CHOICE WITHIN CONSIDERATION SET

Automotive Study	***Delayed Validation***
Decompositional Methods	
HB Logit, Additive Utility	**0.204***
HB Logit, *q*-Compensatory	**0.151***
Direct-Elicitation Methods	
Structured direct-elicitation (Casemap)	0.108
Unstructured direct-elicitation (e-mail)	**0.150***

* Best or not significantly different than best at the 0.05 level.

WEB APPENDIX 2

BRIEF SUMMARY OF DECOMPOSITIONAL METHODS

HB Logit, Additive Utility. Respondents consider a profile if the sum of the partworths of the levels of the profile, plus error, is above a threshold. Subsuming the threshold in the partworth scaling, we get a standard logit likelihood function. We impose a first-stage prior on the partworth vector that is normally distributed with mean $\vec{\beta}_0$ and covariance *D*. The second stage prior on *D* is inverse-Wishart with parameters equal to $I/(N+3)$ and $N+3$, where *N* is the number of parameters to be estimated and *I* is an identity matrix. We use diffuse priors on $\vec{\beta}_0$. Inference is based on a Monte Carlo Markov chain with 20,000 iterations, the first 10,000 of which are used for burn-in.

HB Logit, q-Compensatory. Same as the above except we use rejection sampling to enforce constraints that no feature importance is more than *q* times any other feature importance. We follow Yee, et al. and use $q = 4$, but obtain similar results for $q = 2$, 4, 6, and 8.

Greedoid Dynamic Program. Yee, et al. (2007) demonstrate that a lexicographic ordering of features and levels induces a rank ordering of profiles that has a greedoid structure. This enables us to use forward induction on the feature levels to minimize the number of errors in fitting ordinal paired-comparisons among profiles (vs. observed data) as implied by the feature ordering. The output is a rank ordering of features and levels that best fits the calibration data.

Logical Analysis of Data (LAD). LAD attempts to identify minimal sets of features and levels to distinguish "positive" events from "negative" events (Boros, et. al. 1997). LAD uses a greedy algorithm to find the fewest conjunctive patterns (feature-level combinations) necessary to match the set of considered profiles. The union of these patterns is a disjunction of conjunctions – a generalization of conjunctive, disjunctive, and subset conjunctive decision rules (Gilbride and Allenby 2004, 2006; Jedidi and Kohli 2005). For each respondent, we resolve ties among patterns based on the the frequency of patterns in the sample of respondents. We enforce

cognitive simplicity by limiting the number of feature-levels in a pattern (Hauser, et al. 2010).

WEB APPENDIX 3

DETAILS OF KULLBACK-LEIBLER DIVERGENCE FOR OUR DATA

The Kullback-Leibler divergence (KL) is an information-theory-based measure of the divergence from one probability distribution to another. Because it is calculated for each respondent, we suppress the respondent subscript. We seek the divergence from the predicted consideration probabilities to those that are observed in the validation data , recognizing the discrete nature of the data ($\vec{y}$ such that $y_k = 1$ if the respondent considers profile *k*, 0 otherwise). We predict whether the respondent considers profile *k*. Call this prediction r_k. Let $\vec{r}$ be the vector of the r_k's. If the r_k's were always probabilities (and the number of profiles is not too large), the divergence from the data ($\vec{y}$) to the model being tested ($\vec{r}$) would be:

$$\text{(W1)} \qquad KL = D_{KL}(\vec{y}||\vec{r}) = \sum_{k \in validation} \left[y_k \log_2\left(\frac{y_k}{r_k}\right) + (1 - y_k) \log_2\left(\frac{1 - y_k}{1 - r_k}\right) \right]$$

Equation W1 is poorly defined for discrete predictions (r_k = 0 or 1) and very sensitive to false predictions when r_k approaches 0 or 1. For a fair comparison of both discrete and probabilistic predictions we focus on false positives, true positives, false negatives, and true negatives to separate the summation into four components.[1] Let *V* = the number of profiles in the validation sample, $\hat{C}_v$ = the number of considered validation profiles, F_p = the false positive predictions, and F_n = the false negative predictions. Then the KL divergence is given by the following equation where $S_{c,c}$ is the set of profiles that are considered in the calibration data and considered in the validation data. The sets $S_{c,nc}$, $S_{nc,c}$, and $S_{nc,nc}$ are defined similarly (*nc* → not considered).

$$KL = \sum_{S_{c,c}} \log_2\left(\frac{\hat{C}_v}{\hat{C}_v - F_p}\right) + \sum_{S_{c,nc}} \log_2\left(\frac{V - \hat{C}_v}{F_n}\right) + \sum_{S_{nc,c}} \log_2\left(\frac{\hat{C}_v}{F_p}\right) + \sum_{S_{nc,nc}} \log_2\left(\frac{V - \hat{C}_v}{V - \hat{C}_v - F_n}\right)$$

[1] As per information theory, some information is lost in aggregation. If future researchers develop UDE methods that produce probabilistic predictions, and if the number of profiles is not too large, then comparisons might be made with Equation W1. When comparing discrete and probabilistic predictions we chose to use Equation W2.

After algebraic simplification, KL divergence can be written as:

$$\text{(W2)} \quad KL = D_{KL}(\vec{y}||\vec{r}) = \hat{C}_v \log_2 \hat{C}_v + \left(V - \hat{C}_v\right)\log_2\left(V - \hat{C}_v\right) - \left(\hat{C}_v - F_p\right)\log_2\left(\hat{C}_v - F_p\right) - F_n \log_2 F_n - F_p \log_2 F_p - \left(V - \hat{C}_v - F_n\right)\log_2\left(V - \hat{C}_v - F_n\right)$$

When necessary we use L'hôspital's rule to show that $\lim_{q\to 0} q \log_2 q = 0$.

In the paper we rescale the KL divergence relative to a random null model, specifically: $[D_{KL}(data \parallel random) - D_{KL}(data \parallel model)]/D_{KL}(data \parallel random)$. This scaling is purely for interpretation and does not change the results of any of the statistical tests in this paper.

Equation W2 is related to, but not identical to, the KL measure used by Hauser, et al. (2010), who use the ratio $D_{KL}(model \parallel random)/D_{KL}(data \parallel random)$. Each measure has its own strengths. If we were to use their measure, the basic conclusions would not change. For example, UDE remains significantly better than both Casemap and decomposition for the automotive data ($p < 0.001$). Training effects are similar: UDE improves significantly with training ($p < 0.001$), but Casemap and decomposition do not ($p > 0.05$), UDE is significantly better than Casemap and decomposition with training ($p < 0.001$), and UDE is not significantly different without training ($p > 0.05$). Hit rate and other diagnostic measures reinforce the interpretations that are based on KL.

WEB APPENDIX 4

TASK EVALUATIONS

Mobile Phone Study

We asked respondents whether they understood the tasks and understood that it was "in their best interests to tell us their true preferences." The mean responses on understanding the task were 1.96 (SD = 0.58) and 2.05 (SD = 0.69) for the decompositional and direct-elicitation tasks, respectively, where 1 = "extremely easy", 2 = "easy," 3 = "after putting in effort," 4 = "difficult", and 5 = "extremely difficult. The mean responses for understanding incentive align-

ment were 1.97 (SD = 0.64) and 2.03 (SD = 0.72), respectively. There were no significant differences between the two tasks.

Automotive Study

The mean responses on understanding the task were 1.93 (SD = 0.87), 1.75 (SD=0.75), and 2.34 (SD = 0.99) for the decompositional, Casemap and UDE tasks, respectively, where 1 = "extremely easy," 2 = "easy," 3 = "after putting in effort," 4 = "difficult", and 5 = "extremely difficult." The mean responses for understanding incentive alignment were 1.86 (SD = 0.86), 1.73 (SD=0.80), and 1.89 (SD = 0.88), respectively. Although, the task and the incentives were easiest to understand for Casemap ($p < 0.05$), they appear to be easy to understand for all three methods.

We also asked the participants how the tasks "enable them to accurately express their preferences," where 1 = "very accurately," 3 = "somewhat accurately," and 5 ="not accurately." The mean responses were 2.38 (SD=0.97), 2.15 (SD=0.95), and 2.04 (SD=0.95) for the decompositional, Casemap, and UDE tasks, respectively. Respondents believed the UDE and Casemap tasks enabled them to express their preferences more accurately than the decompositional task ($p < 0.01$), but there is no significant difference between the UDE and the Casemap tasks.

WEB APPENDIX 5

RULES AND PARTWORTHS BY FEATURE LEVEL, AUTOMOBILES

For automobiles the elimination percentages, the compensatory percentages, and the partworths are face valid. As expected, there are differences between direct elicitation and decomposition. As in the mobile phone study, the decompositional partworths are negatively correlated (-0.34) with direct-elicitation elimination percentages and positively correlated (0.50) with direct-elicitation compensatory percentages. The elimination and compensatory percentages are negatively correlated (-0.23).

WEB APPENDIX 6

SCREENSHOTS OF THE STUDIES

Screenshots from both studies will be made available from the authors. They are not included in this document because they would cause the document to be an extremely large file challenging electronic transmission, storage, and printing.