Incentive-Aligned Direct Elicitation of Decision Rules: An Empirical Test

Min Ding

John Hauser

Songting Dong

Daria Silinskaia

Zhilin Yang

Chenting Su

Steven Gaskin*

January 2009

* Min Ding is 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, jhauser@mit.edu.

Songting Dong is a Ph.D. candidate in Marketing, School of Economics and Management, Tsinghua University, Beijing 100084, dongst@sem.tsinghua.edu.cn.

Daria Silinskaia 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)2784-4644; fax (852) 27889146; mkzyang@cityu.edu.hk, mkctsu@cityu.edu.hk.

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

The authors acknowledge financial support from the Research Grant Council of Hong Kong SAR (9041182, CityU 1454/06H), and constructive comments Michael Braun, Ely Dahan, John Liechty, Young-Hoon Park, Vithala Rao, Kayande Ujwal, and from participants in a seminar given at University of Houston.

Incentive-Aligned Direct Elicitation of Decision Rules: An Empirical Test

Abstract

We investigate the feasibility and predictive ability of incentive-aligned directly-elicited rules for decisions consumers make to consider mobile phones. With incentives to think hard and answer truthfully, tested formats ask respondents to state non-compensatory, compensatory, or mixed rules for an agent who will select a product for the respondents. The direct-elicitation methods are designed to be scalable to large numbers of features and feature levels; designs that are challenging for decompositional methods. Two validation tasks, one at the time of the survey and one delayed three weeks, ask respondents to indicate which of 32 mobile phones they would consider from a fractional $4^5 2^2$ design of features and levels. Predictions are compared to compensatory, non-compensatory, and mixed decompositional methods chosen to represent current practice. Mixed models (decomposition or direct-elicitation) predict better than pure models. For the initial validation task (which shares common-methods effects with the decompositional-estimation task), a mixed decompositional model does better on hit rate, but not on an information-theoretic measure. Decomposition and direct elicitation become comparable when validation is delayed three weeks (and are comparable for choice within the consideration set). We close by illustrating managerial insights and testing scalability.

Keywords: *Decision rules, conjoint analysis, consideration sets, non-compensatory rules, product development, incentive alignment.*

INTRODUCTION AND PROBLEM STATEMENT

Decompositional conjoint analysis has been one of the most successful quantitative research methods in marketing providing insight on consumer preferences for product or service features (e.g., Green 2004; Green, Krieger, and Wind 2001). Recently, behavioral theory and managerial practice have pushed conjoint analysis in three scientifically interesting directions.

First, applications such as automobiles and high-tech gadgets have become rich in features requiring large numbers of profiles for even orthogonal experimental designs. For example, Silinskaia, et al. (2009) describe a modest conjoint design used by a US auto maker to develop strategies by which the auto maker could increase consideration of its vehicles. That study focused on a subset of the auto market with 21 brands, 9 body types, 7 price levels, and five other features each with 5 or fewer levels. Pure decompositional methods would have required a minimal orthogonal design of 13,320 profiles.[1] Respondents cannot evaluate such large numbers of profiles. Practitioners have told us that they seek practical, easy-to-implement methods that scale up to the large numbers of features and feature-levels typical in complex products and services. We seek a method that scales well to such large designs.

Second, in web-based purchasing, catalogs, and superstores consumers often select from among 20 to 100+ products. When faced with so many alternatives, behavioral research suggests that consumers use a two-stage consider-then-choose process rather than a one-stage compensatory evaluation (e.g., Hauser and Wernerfelt 1990; Payne 1973; Roberts and Lattin 1991; Swait and Erdem 2007). Because consideration set sizes are often a small fraction of the options available (e.g., 8 of 300+ automobile brands), it is scientifically interesting and managerially relevant to study the consideration stage. We seek a method that recognizes such two-stage processes

[1]One might use Bayesian shrinkage asking each respondent to evaluate a small fraction of the full design. Accuracy depends upon the relative sparseness of the data. In a later section we demonstrate that our incentive-aligned direct-elicitation methods scale well to this particular design. A full-factorial would have been 357,210 profiles.

when relevant and gathers data that can be either for consideration, choice, or consider-then-choose. (For ease of exposition and other reasons, we focus primarily on the consideration stage relegating the choice stage to a supplemental appendix.)

Third, particularly when faced with many feature-rich products, behavioral research and decompositional methods suggest that some consumers use decision heuristics, such as lexicographic, conjunctive, or disjunctive rules, to balance cognitive costs and decision benefits (Gilbride and Allenby 2004, 2006; Jedidi and Kohli 2005; Kohli and Jedidi 2007; Payne, Bettman and Johnson 1988, 1993; Yee, et al. 2007). Methods based on both compensatory and non-compensatory decision rules have the potential to be more accurate when predicting future consideration and choice and, perhaps, provide more managerially useful descriptions of consumers' decision rules.

Direct elicitation of consumers' decision rules is one method that might scale to large number of features and levels, focus on consideration and/or choice, and allow both compensatory and non-compensatory decision rules. However, it is a valid concern that consumers may not have incentives (or may not be able) to describe their own decision rules. Recent developments in preference measurement justify a fresh look at direct-elicitation methods. (1) Improvements in compensatory-model direct elicitation have increased accuracy in many contexts (Netzer and Srinivasan 2007; Park, Ding, and Rao 2008; Srinivasan 1988; Srinivasan and Wyner 1998). Indeed, one of the largest and most quoted conjoint designs in the academic literature used direct elicitation as part of hybrid data collection (Wind, et al. 1989). (2) Incentive-aligned designs, which involve natural tasks that consumers do in their daily life with real consequences, lead to greater respondent involvement, less boredom, and higher data quality. Research suggests that incentive-aligned decompositional methods enhance predictions and face-validity

(Ding 2007; Ding, Grewal and Liechty 2005). (3) The diffusion of "voice-of-the-customer" methods has created practical expertise within many market research firms in the cost-effective quantitative coding of qualitative data (e.g., Griffin and Hauser 1993; Perreault and Leigh 1989). These firms have skills that are well-matched to the field application of incentive-aligned direct-elicitation methods. If incentive-aligned direct-elicitation proves feasible and accurate this expertise implies such methods might diffuse rapidly.

Motivated by these recent developments we investigate the feasibility and relative accuracy of scalable direct-elicitation methods. We adopt incentive alignment as a best practice to focus on the comparison of direct elicitation to decomposition. We are particularly interested in consideration decisions and elicitation procedures that allow compensatory, non-compensatory, or mixed decision rules. We chose a primary context, mobile phones in Hong Kong, in which both direct elicitation and decompositional approaches are feasible. The mobile phone products and features are sufficiently familiar to respondents that compensatory rules are feasible; at the same time the features have sufficiently many levels that respondents might choose to use non-compensatory rules. This context provides a reasonable test of incentive-aligned direct elicitation as compared to incentive-aligned compensatory, non-compensatory, and mixed decompositional methods. After comparing relative predictive accuracy in a situation in which <u>both</u> direct elicitation and decomposition are feasible, we demonstrate that direct elicitation scales well to experimental designs where decomposition is not feasible (or at least very challenging) at the level of the individual respondent.

Our research goals are "proof of concept" and "initial test." We seek to demonstrate that an incentive-aligned direct-elicitation method can be designed to match or beat the most commonly-used state-of-the-art decompositional methods in situations were decompositional methods are feasible. With time and subsequent research, others might apply incentive-alignment, consideration (then choice), and compensatory/non-compensatory modeling to other direct-elicitation methods. Likewise, we choose decompositional benchmarks that use a variety of methods and which have done well in previous comparative testing. By investigating the "pure" forms of direct-elicitation and decomposition we also gain insight to inform tradeoffs in hybrid designs.

PREVIOUS LITERATURE

Direct elicitation (sometimes called self explication) has been used to measure consumer preferences and/or attitudes for over forty years either alone or in combination with decompositional methods (Fishbein 1967; Fishbein and Ajzen 1975; Green 1984; Sawtooth 1996; Hoepfl and Huber 1975; Wilkie and Pessemier 1973). The accuracy of direct elicitation of compensatory rules has varied considerably relative to decompositional methods (e.g., Akaah and Korgaonkar 1983; Bateson, Reibstein and Boulding 1987; Green 1984; Green and Helsen 1989; Huber, et al. 1993; Leigh, MacKay and Summers 1984, Moore and Semenik 1988; Srinivasan and Park 1997). Attempts at the direct elicitation of non-compensatory rules have met with less success partly because respondents often choose profiles with levels they say are "unacceptable" (Green, Krieger and Banal 1988; Klein 1986; Srinivasan and Wyner 1988; Sawtooth 1996).

Recent applications of incentive-aligned compensatory conjoint analysis provide examples where respondents are encouraged to provide truthful and thoughtful answers because the rewards they receive are based on the answers they give to questions (Ding 2007; Ding, Grewal and Liechty 2005). All tests to date enhance predictive ability. Such incentive-aligned conjoint analysis draws on the precepts of experimental economics and is becoming common in other market-research applications (Smith 1976; Prelec 2004). Incentive-aligned decompositional conjoint-analysis applications include laptops, wine, and laptop bags (Kugelberg 2004; Toubia, Hauser and Garcia 2007; Toubia, et al. 2003). Other than Park, Ding and Rao (2008), we are unaware of any direct-elicitation conjoint-analysis methods (compensatory or non-compensatory) that use incentive-aligned methods.

Decompositional methods have been proposed for conjunctive, disjunctive, subset conjunctive, lexicographic, and disjunctions of conjunctions (Gilbride and Allenby 2004, 2006; Hauser, et al. 2008; Jedidi and Kohli 2005; Kohli and Jedidi 2007; Moore and Karniouchina 2006; Yee, et al. 2007).[2] Results to date suggest that non-compensatory methods predict comparably to, but sometimes less well than, compensatory methods in product categories with which respondents are familiar (batteries, computers). Non-compensatory methods are slightly better in unfamiliar categories (smartphones, GPSs). Research suggests that approximately one-half to two-thirds of the respondents are fit better with non-compensatory rather than compensatory methods and that the percentage is higher when respondents are asked to evaluate more profiles. The results are comparable whether the decision is consideration, consider-then-choose, or choice. We are unaware of any comparisons to non-compensatory direct-elicitation methods.

PRODUCT CATEGORY, SUBJECTS, AND STUDY DESIGN

In Hong Kong, mobile phone shops line every street with "an untold selection of manufacturers and models (German 2007)." "The entire [mobile] phone culture is far advanced" with consumers able to buy unlocked mobile phones that can be used with any carrier (ibid.). Using local informants, observation of mobile phone stores, and discussions with potential respondents

[2] A conjunctive rule eliminates profiles with features that are not above minimum levels. A disjunctive rule accepts a profile if at least one feature is above a defined level. Subset conjunctive rules require that *S* features be above a minimum level. Disjunction of conjunctions rules generalize these rules further. A profile is acceptable if its features are above minimum levels on one or more defined subsets of features. Lexicographic rules order features. The feature ordering implies a profile ordering based on the highest ranked feature on which the profiles vary. For consideration decisions, a lexicographic rule degenerates to a conjunctive model with an externally-defined cutoff.

we selected a set of features and feature-levels that represent the choices faced by Hong Kong respondents. Pretests indicated the following feature-levels were face valid:

- Brand: Motorola, Lenovo, Nokia, Sony-Ericsson
- Color: black, blue, silver, pink
- Screen size: small (1.8 inch), large (3.0 inch)
- Thickness: slim (9 mm), normal (17 mm)
- Camera resolution: 0.5 Mp, 1.0 Mp, 2.0 Mp, 3.0 Mp
- Style: bar, flip, slide, rotational
- Base price level: $HK1080, $HK1280, $HK1480, $HK1680 [$1 ≈ $HK8]

To make the respondent's task realistic and to avoid dominated profiles (e.g., Elrod, Louviere and Davey 1992; Johnson, Meyer and Ghose 1989), the price levels for each profile were the sum of an experimentally-varied base price level plus an increment for relevant feature-levels (e.g., if a profile has a large screen we add $HK200 to the price). The resulting profile prices ranged from $HK1080 to $HK2480. Prior research suggests that such Pareto designs do not affect predictability substantially nor inhibit the non-compensatory use of price (Green, Helsen, and Shandler 1988; Hauser, et. al. 2008; Toubia, et al. 2003).

This $4^5 2^2$ design is typical of compensatory decompositional conjoint analysis and at the upper limit of non-compensatory decompositional methods which require computations that are exponential in the number of feature levels. Larger designs might not be feasible for non-compensatory decompositional methods. The web-based survey was pretested with 56 respondents. Pretests indicated that the questions were clear and the task not onerous.

The subjects were students at a major university in Hong Kong who were screened to be 18 years or older and interested in purchasing a mobile phone. All subjects came to a computer laboratory on campus to complete the web-based survey, and then completed a delayed validation task on any internet-connected computer 3 weeks later. Those who completed both tasks received $HK100 and were eligible to receive an incentive-aligned prize (as described below). In total 143 respondents completed the entire study and provided data with which to estimate the decision rules. This represents a completion rate of 88.3%.

We focus on the consideration task where, for the purpose of this study, consideration was defined to respondents as "mobile phones that you would purchase if your most-preferred mobile phone were unavailable." While other definitions of "consideration" exist in the marketing literature, pretests indicated that this definition was most natural to way that Hong Kong subjects evaluated and chose mobile phones. We focus on the consideration task rather than the choice task (1) because of growing managerial and scientific interest in consideration decisions, (2) direct elicitation of consideration rules is relatively novel in the literature, (3) the consideration task was more likely to provide a test of compensatory, non-compensatory, and mixed decision rules, (4) the predictive ability for the choice task (rank order within the consideration set) mimics the basic results we obtain for the consideration task, and (5) focused exposition. (Choice-task results are summarized in a supplemental appendix available from the authors.)

To obtain greater statistical power we use a within-subjects design in which subjects complete both a direct-elicitation and a decompositional task. We use two validation tasks. One task occurs toward end of the web-based survey after a memory-cleansing task and a second task is delayed by three weeks. The validation tasks use an interface identical to the decompositional task thus ensuring that common-methods effects favor the decompositional tasks relative to the direct-elicitation task. For ease of exposition, we call the first decompositional task the *Calibration Task,* the first validation task the *Initial Validation Task,* and the second validation task the *Delayed Validation Task*. Specifically, the survey proceeded as follows:

- Initial screens assured privacy and described the basic study.
- Mobile-phone features were introduced one feature at a time through text and pictures.

- Incentives were described for both the decompositional and direct-elicitation tasks.
- The order of the following two tasks was randomized.
  - Respondents indicated which of 32 mobile phones they would consider (*Calibration Task*). Considered profiles were ranked (see supplemental appendix).
  - Respondents described decision rules to be used by an agent to select a mobile phone for the respondent (*Structured Direct-Elicitation Task*, details below).
- "Brain-teaser" distraction questions cleared short-term memory (Frederick 2005).
- Respondents saw a new set of 32 mobile phones and indicated those they would consider (*Initial Validation Task*). Considered profiles were ranked (see appendix).
- Respondents were asked to write an e-mail as an alternative way to instruct an agent to select a mobile phone (*E-mail-based Direct-Elicitation Task,* details below).
- Short questions measured respondents' comprehension of the incentives and tasks.
- (Three weeks later). Respondents saw a third set of 32 mobile phones and indicated those they would consider (*Delayed Validation Task*).

Caveats. This design focuses on methods comparison. At minimum, we believe the study design has internal validity. We chose features to represent the Hong Kong market and we chose the consideration task to represent the typical Hong Kong store. However, the most difficult induction for consideration decisions is the cognitive evaluation cost. If the evaluation cost in the survey varies from the market, the consideration-set size in a real store might differ from the consideration-set size in a survey. Nonetheless, the evaluation cost is constant between methods because the comparison between decompositional and direct-elicitation methods is based on the same validation data (initial and delayed). We hope, but cannot prove, that the incentives also enhance external validity. At minimum, pretest comments and post-survey debriefs suggest respondents believed they will behave in the market as they did in the survey. (In addition, respondents who received mobile phones as part of the incentives were satisfied with the mobile phones that were chosen for them.)

A second concern is that either the decompositional estimation task or the direct-elicitation task trains respondents, perhaps affecting how respondents construct decision rules (e.g., Payne, Bettman and Johnson 1993). If so, this would enhance internal consistency. The delayed task is one attempt to minimize that effect. However, any internal consistency would favor decompositional methods which use the same type of task for estimation as validation. Training tasks are common market-research practice and are not all bad. Many professionals hypothesize that training enhancement increases external validity if it causes respondents to think as carefully about mobile phone purchases in the survey as they would in a $HK1000-2500 purchase in the store. But training enhancement remains an hypothesis to be tested in future studies.

A third concern is an order effect for one of the two direct-elicitation tasks (the e-mail task) which occurs after the initial validation task but three weeks prior to the delayed validation task. The initial design focused on a formal structure for direct elicitation with a goal of simplifying coding. The e-mail task was a second direct-elicitation task which has the advantage of requesting decision rules in a more natural format but the disadvantage that coding requires a judging methodology (e.g., Hughes and Garrett 1990; Perreault and Leigh 1989; Wright 1973). For predictions based on the e-mail task we must interpret the <u>initial</u> validation cautiously. Fortunately potential order effects are mitigated for the <u>delayed</u> validation task.

RESPONDENT TASKS AND INCENTIVES

Decompositional Tasks and Benchmark Models

The decompositional calibration task and the validation tasks displayed three panels following a format developed by Hauser, et al (2008). The left panel showed icons representing the 32 mobile phones. Profiles were chosen from an orthogonal fractional factorial of the $4^5 2^2$ design. When the respondent clicked on an icon, the mobile phone appeared in the center panel (features were described by pictures and text). The respondent indicated whether or not he or she would consider that mobile phone. Considered phones appeared in the right panel. The respondent could reverse the panel to see not-considered phones and could move phones among considered, not-considered, and to-be-evaluated until the respondent was satisfied with his/her consideration set. The data to estimate the decompositional models are 0-vs.-1 indicators of whether each profile is included in the consideration set or not.

We chose as benchmarks commonly used compensatory and non-compensatory decompositional methods. Our first benchmark is the standard hierarchical Bayes logit model applied to consideration sets using the 32 consider-vs.-not-consider observations per respondent (Hauser, et al. 2008; Lenk, et al. 1996; Rossi and Allenby 2003, Sawtooth 2004; Swait and Erdem 2007). The specification is an additive partworth model. Many researchers have argued that compensatory models, lexicographic models, subset conjunctive, and conjunctive models can be represented by such an additive partworth model (e.g., Jedidi and Kohli 2005; Kohli and Jedidi 2007; Olshavsky and Acito 1980; Yee, et al. 2007).[3] Following Bröder (2000) and Yee, et al. (2007) we also specify a *q*-compensatory model by constraining the additive model so that no feature's importance is more than *q* times as large as another feature's importance. (A feature's importance is the difference between the maximum and minimum partworths for that feature.) The *q*-compensatory model limits decision rules so that they are compensatory; the unconstrained additive-partworth model is consistent with both compensatory and non-compensatory decision rules.

There are a variety of non-compensatory decompositional models we can choose as benchmarks. We select two that have done well in previous research. The first is the greedoid

[3]Examples: If there are F feature levels and if the partworths are, in order of largest to smallest, 2^{F-1}, 2^{F-2}, …, 2, 1, then the additive model will act as if it were lexicographic by aspects. If S partworths have a value of β, the remaining partworths a value of 0, and if the utility cutoff is $S\beta$, then the model will act as if it were conjunctive. The analytic proofs assume no measurement error.

dynamic program which estimates a lexicographic consideration-set model (Yee, et al. 2007). The second is logical analysis of data which has matched or outperformed other non-compensatory estimation methods, including hierarchical Bayes specifications of conjunctive, disjunctive, and subset conjunctive models, in at least one study (Gilbride and Allenby 2004, 2006; Jedidi and Kohli 2005; Hauser, et al. 2008). Logical analysis of data estimates disjunctions-of-conjunctions rules, which are generalizations of disjunctive, conjunctive, subset conjunctive, and in the case of consideration data, lexicographic rules (Boros, et. al. 1997; 2000). We hope that together these two models, which themselves have been compared to a variety of other models and which represent different perspectives, provide reasonable initial benchmarks to represent a broader set of non-compensatory decompositional models. (A Supplemental Appendix, available from the authors, summarizes the benchmark methods.)

Direct-Elicitation Tasks

A structured direct-elicitation task asked respondents to provide rules for a friend who would act as their agent in purchasing a mobile phone for them. Respondents were asked to state instructions unambiguously and to state as many instructions as were necessary. The initial screen had open boxes for five rules. Respondents were not required to state five rules and they could add rules if desired. An unstructured direct-elicitation task asked respondents to state their instructions to the agent in the form of an e-mail to a friend. Other than a requirement to start the e-mail with “Dear friend,” respondents could use any format to describe their decision rules.

The two direct-elicitation tasks were coded by two independent judges who were blind to any hypotheses. After coding independently, the two judges met to reconcile differences. The judges were extremely consistent in extracting rules from direct-elicitation tasks. Such coding is common in market research for both commercial use and for litigation (e.g., Hughes and Garrett

1990; Perreault and Leigh 1989; Wright 1973). The coding guide, the transcripts, and all coded responses are available from the authors.

Explicit elimination rules were coded as such (-1 in the database) and used to eliminate profiles in any predictions of consideration. Acceptance rules, such as “only buy Nokia,” imply that all brands but Nokia are eliminated. Compensatory preferences were assigned an ordinal scale. For example, if the respondent says he or she prefers Nokia, Motorola, Lenovo, and Sony-Ericsson in that order (and does not eliminate any brand), then Nokia would be assigned a “1,” Motorola a “2,” Lenovo a “3,” and Sony-Ericsson a “4.” In predictions these ratings are treated as ordinal ratings. We did not attempt to code the relative preferences among different features. This results in weak orders of profiles (ties allowed) and is thus conservative. We chose this conservative coding strategy so that predictions were not overly dependent on our judges’ subjective judgments and so that their judgments would be more readily reproduced.

To illustrate the coding we provide example statements from respondents’ e-mails (retaining original language and grammar). Based on the judges’ classifications of these statements, over three-fourths of the respondents (78.3%) asked their agents to use a mixture of compensatory and non-compensatory rules for consideration and/or choice. Most of the remainder were compensatory (21.0%) and only a very few were purely non-compensatory (0.7%).

> **(Mostly non-compensatory)**. *Dear friend, Please help me to buy a mobile phone. And there are some requirements for you to select it for me: 1. Camera better with 3.0mp, but at least 2.0 2. Only silver or black 3. Only select Sony Ericsson or Nokia. Thank you for your help.* [Coding: -1 for 1.0 Mp, 0.5 Mp, Motorola, Lenovo, blue, and pink. 1 for 3.0 Mp.]

> **(Mixed non-compensatory/compensatory)**. *Dear friend, I want to buy a mobile phone recently …. The following are some requirement of my preferences. Firstly, my budget is about $2000, the price should not more than it. The brand of mobile phone is better Nokia, Sony-Ericsson, Motorola, because I don't like much about Lenovo. I don't like any mobile phone in*

pink color. Also, the mobile phone should be large in screen size, but the thickness is not very important for me. Also, the camera resolution is not important too, because i don't always take photo, but it should be at least 1.0Mp. Furthermore, I prefer slide and rotational phone design. It is hoped that you can help me to choose a mobile phone suitable for me. [Coding: -1 for 0.5 Mp, pink, small screen, 1 for slide and rotational, and 4 for Lenovo. Our coding is conservative. For this respondent, neither the subjective statements of relative importances of features nor the target price were judged sufficiently unambiguous to be coded.]

(Mostly compensatory). *Dear friend, I would like you to help me buy a mobile phone. Nokia is the most favorite brand I like, but Sony Ericsson is also okay for me. Bar phones give me a feeling of easy-to-use, so I prefer to have a new bar phone. The main features which I hope to be included in the new mobile phone are as follows: A: 2Mp camera resolution B: Black or Blue color C: Slimness in medium-level D: Pretty large screen Hopefully my requirements for the purchase of this mobile phone are not too demanding, thank you for you in advance.* [Coding: 1 for Nokia, bar, 2.0 Mp, black, blue, small size, large screen, and 2 for Sony Ericsson. The respondent's statement ranks 2.0 Mp above 3.0 Mp, which is consistent with the market and our design because 3.0 Mp is priced higher.]

The coded responses provide a partial order for any set of mobile phones. Elimination or acceptance rules exclude or include some mobile-phone profiles. Preference rules provide a weak ordering (ties allowed) of non-excluded/included profiles.

Incentives

Designing aligned incentives for the consideration task is challenging because consideration is an intermediate stage in the decision process. Other researchers have used purposefully vague statements that were pretested to encourage respondents to trust that agents would act in the respondents' best interests (e.g., Kugelberg 2000). For example, if we told respondents they would get every mobile phone considered, the best response is a large consideration set. If we told respondents they would receive their most-preferred mobile phone, the best response is a consideration set of exactly one mobile phone. Instead, based on pretests, we chose the following two-stage mechanism. Because this mechanism is an heuristic, we call it "incentive aligned" rather than the more-formal term, "incentive compatible." Our goals with incentive alignment are: (1) the respondents believe it is in their interests to think hard and tell the truth, (2) it is, as much as feasible, in their interests to do so, and (3) there is no way, that is obvious to the respondents, by which they improve their welfare by "cheating."

However, our results must be viewed as a conservative initial test (for direct elicitation). Any misalignment in the incentives favors decompositional methods relative to direct-elicitation methods because the tasks we use to validate predictions share incentive-alignment issues with the task used to estimate the parameters of the decompositional models.

Specifically, respondents were told they had a 1-in-30 chance of receiving a mobile phone plus cash representing the difference between the price of the phone and $HK2500.[4] Because we wanted both the direct-elicitation and decompositional tasks to be incentive-aligned, respondents were told that one of the tasks would be selected by "coin flips" to determine their prize. In addition, respondents were reminded: "It is in your best interest to think carefully when you respond to these tasks. Otherwise you might end up with something you prefer less, should you be selected as the winner."

For the decompositional task respondents were told we would first randomly select one of the three tasks (2 in the main study and 1 in the delayed study), and then select a random subset of the 32 phones in that task. Respondents' consideration decisions in the chosen task would determine which phone they received. If there were more than one phone that matched their

[4] The prize of $HK2500, approximately $US300+, might induce a wealth-endowment effect making the respondent more likely to choose more features. While this is an interesting research opportunity, a priori it should not favor decomposition over direct elicitation or vice versa. In one example with decompositional methods, Toubia, et al. (2003) endowed their respondents with $100. They report good external validity when forecasting market shares after the product was launched to the market.

consideration set the rank data would distinguish the phones. (Model-comparison results on the rank data mimic those on consideration data. See supplemental appendix.) The unknown random subset is important here. If respondents "consider" too many or too few profiles they may not receive an acceptable mobile phone should they win the lottery. The incentives are aligned for both consideration (our primary focus) and choice (supplemental appendix).

For the direct-elicitation tasks respondents were told that two agents would use the respondents' decision rules to select a phone from a secret list of mobile phones. If the two agents disagreed, a third agent would settle the disagreement. To encourage respondents to trust the agents, respondents were told the agents would be audited and not paid unless the respondents' instructions were followed accurately (e.g., Toubia 2006).

At the conclusion of the study, five respondents were selected randomly, and each received a specific mobile phone (and cash) based on the mechanism described above. All respondents received the fixed participation fee (HK$100) as promised.

To examine the face validity of the incentive alignment, 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 compatibility were 1.97 (SD = 0.64) and 2.03 (SD = 0.72), respectively. There were no significant differences between the two tasks. Basically, on average, respondents found the tasks and incentive alignment easy to understand. Qualitative statements also suggested that respondents believed that their answers should be truthful and reflect their true consideration decisions.

Results

Descriptive Statistics

The average size of the consideration set was 9.3 in the (decompositional) Calibration Task. Consideration-set sizes were comparable for the Initial Validation Task (9.4) and the Delayed Validation Task (9.3). All are statistically equivalent, consistent with an hypothesis that respondents thought carefully about the task.

Ability to Predict Consideration Sets in the Validation Tasks

Comparative statistics. Hit rate is an intuitive measure with which to compare predictive ability. However, hit rate must be interpreted with caution for consideration data. With average consideration sets around 9.3 out of 32 (29.1%), a null model that predicts that no mobile phones will be considered will achieve a hit rate of 70.9%. The apparent strong performance of such a null model suggests that we gain insight by reporting other statistics. For example, a good model should get reasonably close when predicting the consideration-set size. But a null model of random prediction (proportional to consideration-set size) gets the consideration-set size correct but achieves a hit rate of 58.3%. Thus, we follow Srinivasan (1988), Srinivasan and Park (1997), and Payne, Bettman and Johnson (1993, p. 128) and report the percent improvement relative to a random-prediction null model. We also report the predicted consideration-set sizes.

One problem with either hit-rate measure is that they merge false positives and false negatives. If we knew the managerial situation we might specify loss functions to weigh false positives and false negatives differently. Alternatively, we use the Kullback-Leibler divergence (K-L) which measures the expected gain in Shannon's information measure relative to a random model (Chaloner and Verdinelli 1995; Kullback and Leibler 1951). The K-L percentage is 0% for both the random null model and the consider-all-profiles null model. It is 100% for perfect prediction. The K-L percentage rewards models that predict the consideration-set size correctly and favors a mix of false positives and false negatives that reflect true consideration sets over those that do not.[5] It discriminates among models even when the hit rates might otherwise be equal. Together the four statistics, hit rate, relative hit rate improvement, predicted consideration-set size, and the K-L percentage, provide a means to assess relative predictive ability.

Predicting with directly-elicited rules. Most respondent statements used mixed non-compensatory/compensatory rules (78.3%). However, for diagnostic purposes it is also useful to examine the predictive ability of the elimination rules without the compensatory component. Thus, we examine four perspectives on the directly-elicited rules: two elimination-only models and two mixed models.

The first elimination-only model, which we label "Raw Conjunctive," uses only the explicit non-compensatory statements. Detailed observation of the qualitative data suggest that the Hong Kong respondents were extremely polite and often stated implicit elimination rules by mentioning preference for more than half of the levels for a feature. For example, rather than eliminating pink phones the respondent might express a preference for blue, black, or silver. Thus, our second elimination-only model, which we call "Modified Conjunctive," includes as elimination rules such recoded responses. We implemented the recoding as mechanical rules. The coding guide was developed based on a two-thirds subset of the Calibration Task data and tested on the remaining Calibration Task data. No validation data were used to construct the coding guide.

In our two mixed models, we complement the explicit conjunctive rules with the compensatory statements that weakly order non-eliminated profiles. The order is weak because the qualitative statements may not distinguish tradeoffs among features or levels within features

[5] Formulae for K-L percentage for consideration-set prediction are available in Hauser, et al. (2008). K-L acts for 0-vs.-1 predictions much like U^2 does for probabilistic predictions (Hauser 1978). Example calculations to support the claims in this paragraph are available from the authors.

(e.g., “I prefer phones that are black or silver and flip or slide.”). To predict a consideration set with such compensatory statements we need to establish a utility threshold that balances benefits of a larger consideration set with the cognitive costs. We do this in two ways. “Mixed, Match Cutoff” selects a threshold so that the predicted consideration-set size matches, as nearly as feasible, the consideration-set size in the estimation data. The match is not perfect because weak preference orders make the threshold slightly ambiguous.

Using calibration consideration-set sizes favors neither decompositional nor direct-elicitation methods because the threshold is also implicit in all of the decompositional estimation methods. However, to be conservative, we also test a mixed model which estimates the consideration-set-size threshold using a binary logit model with the following explanatory variables: the stated price range, the number of non-price elimination rules, and the number of non-price preference rules. We label this model “Mixed, Logit-Based Cutoff.”

[Table 1 about here.]

Comparisons. The unstructured e-mail direct-elicitation task does significantly better than the structured task in all eight hit-rate comparisons (4 direct-elicitation models x 2 validation tasks), and significantly better on seven of the eight K-L percentage comparisons. Thus, for ease of exposition, we focus on the e-mail-based direct-elicitation methods. Detailed comparisons between the structured and unstructured tasks are available from the authors. Table 1 summarizes the remaining predictive tests.

We focus first on the Initial Validation Task. The best decompositional method is the “HB Logit with Additive Utility.” The best direct-elicitation method is “Mixed, Match Cutoff.” Based on the qualitative observation that most directly-elicited statements contain both compensatory and non-compensatory instructions, it is not surprising that the mixed models do well.

Comparing decompositional methods to direct-elicitation methods we see that "HB Logit, Additive Utility" is significantly better than the best direct-elicitation on the hit-rate measures, but not on the K-L percentage. In contrast, the mixed e-mail direct-elicitation methods do slightly better on K-L percentage, although not significantly so.

All mixed-rule methods do reasonably well in predicting consideration-set size – all are within roughly a single profile (e.g., 8.6 vs. 9.3 profiles).[6] Results for non-compensatory-only methods are mixed. Decompositional methods do well but elimination-only direct-elicitation methods do substantially less well. This, too, is not surprising. Decompositional methods select rules to fit calibration data. The algorithms might choose non-compensatory rules which approximate mixed rules. Finally, the pure compensatory method ("HB Logit, *q*-Compensatory") does not do as well as the mixed-rule methods.

We next focus on the Delayed Validation Task. We see similar patterns except that a direct-elicitation method ("Mixed, Match Cutoff") is now statistically comparable to "HB Logit with Additive Utility." Perhaps the three-week delay mitigated common-measurement effects between the decompositional estimation task and the validation task.

Based on Table 1 we tentatively conclude that an unstructured e-mail direct-elicitation method with estimated thresholds gives comparable predictions to the additive decompositional method, especially for the Delayed Validation Task. Predictions are consistent with the qualitative interpretation that most respondents use mixed non-compensatory/compensatory decision rules. We believe these are important findings. While direct-elicitation tasks are comparable in difficulty to decomposition for moderate numbers of features and feature-levels, direct elicitation is more likely to scale to practical applications that require large numbers of features and feature-

[6] Although the goal of "Mixed, Match Cutoff" is to match the consideration-set size, this model slightly under predicts the consideration-set sizes due to ambiguity (ties) in the preference rules. This ambiguity might be resolved if we choose to be less conservative in the coding guide for the compensatory rules.

levels. Also, the mixed incentive-aligned direct-elicitation methods are comparatively new relative to incentive-aligned decomposition and may improve with further application.

Other comments. The decompositional non-compensatory methods are surprisingly robust. While the mixed (additive) decompositional model predicts hit rates better than either the non-compensatory models or the *q*-compensatory model, the improvement in hit rate is not as overwhelming as it is for "raw conjunctive."[7] On K-L percentage the decompositional non-compensatory models are comparable to the additive model and superior to the *q*-compensatory model, especially on delayed validation. This predictive performance is consistent with results in Yee, et al. 2007. (We follow Yee, et al. and use $q = 4$, but obtain similar results for $q = 2, 4, 6$, and 8.)

Choice within the consideration set. The methods' relative abilities to predict rank-order preference <u>within the consideration set</u> are consistent with the methods' relative abilities to predict consideration. The best predictions are obtained with the mixed-rule direct-elicitation method and the additive HB logit method. Neither are statistically different from one another and both are statistically better than the pure compensatory and the pure non-compensatory methods. Details are in the Supplemental Appendix.

MANAGERIAL OUTPUTS

The managerial presentation of decompositional additive partworths has been developed through decades of application. The growing popularity of HB adds heterogeneity with corresponding challenges in how best to describe heterogeneity to managers. The last two columns of Table 2 provide a commonly-used format – the posterior means and standard deviations (across respondents). For example, the pink color has, on average, a large negative partworth, however,

[7] One might wonder about the moderately high K-L percentage for "raw conjunctive" despite the mismatch in <u>average</u> consideration-set size. The explanation is bimodality. The "raw conjunctive" model does extremely well matching consideration set sizes on some respondents. The average K-L for the remaining respondents is much lower.

not all respondents agree because the heterogeneity among respondents is large. Indeed, with few exceptions the "HB Logit, Additive Utility" model suggests that our respondents vary considerably in their preferences for mobile-phone features.

Applications of non-compensatory preference measurement are relatively recent. Academics and practitioners are still evolving the best way to summarize non-compensatory preferences for managers. Table 2 provides one potential summary. The third column reports the percent of respondents whose directly-elicited decision rules include a feature level as an elimination criterion. For example, 12.6% of the respondents mention that they would eliminate any Motorola mobile phone whereas only 1.4% would eliminate any Nokia mobile phone. The highest non-compensatory feature levels are low camera resolutions and the pink color. For example, 29.4% of the respondents would not buy pink mobile phones. Price is treated slightly differently from other features in our design because the prices that respondents saw were a combination of the base price manipulation and feature-based increments. Nonetheless, 18.9% of the respondents stated they would only accept mobile phones within specific price ranges.

[Table 2 about here.]

We attempt to summarize respondents' directly-elicited compensatory statements in the fourth column of Table 2 by displaying the percent of respondents who mentioned each of the feature levels in a compensatory rule. (Respondents might mention one feature level, multiple feature levels, or none at all.) For example, more than half (60.1%) of our respondents mentioned Nokia. High camera resolutions were also mentioned by large percentages of respondents. Interesting, the percent of compensatory mentions from direct elicitation is significantly correlated with the "HB Logit, Additive Utility" posterior mean partworths ($\rho = 0.72$, $p < 0.001$). As befits a mixed model, the posterior means of the partworths are also significantly negatively correlated with the directly-elicited feature-elimination percentages ($\rho = -0.49$, $p < 0.02$).

We have no prediction with respect to the correlation between directly-elicited non-compensatory and compensatory percentages. In our application the directly-elicited compensatory percentages are significantly negatively correlated with directly-elicited elimination percentages ($\rho = -0.71$, $p < 0.001$). For example, camera resolution affects many respondents' decisions, but for some respondents low resolution is an elimination feature and for other respondents high resolution is an important, but compensatory feature.

There are other ways we might summarize the output of non-compensatory/compensatory direct-elicitation preference measurement depending upon the specific managerial question being addressed. For example, if Lenovo were considering launching a $HK2500, pink, small-screen, thick, rotational phone with a 0.5 Mp camera resolution, the majority of respondents (67.8%) would not even consider it. On the other hand, almost everyone (all but 7.7%) would consider a Nokia, $HK2000, silver, large-screen, slim, slide phone with 3.0 Mp camera resolution. We might also use the respondent-level direct-elicitation data to identify market segments (an analogy to what is now done with respondent-level partworth posterior means).

SCALABILITY

As an initial test of scalability we had the opportunity to piggy-back on a decompositional study funded by a US automaker. We were allowed to include an e-mail direct-elicitation task in a pretest of the survey. Because an orthogonal fraction of the $21 \times 9 \times 7 \times 5 \times 3^3 \times 2$ design was too large for respondents to evaluate, the automaker asked each respondent to evaluate 30 of the 13,320 vehicles in an orthogonal design. After completing the automaker's survey, pretest respondents completed an additional survey to evaluate the feasibility of the direct-elicitation task.[8]

[8] At present, we have access only to data from the second (feasibility) survey. The full-scale study will test inductions to increase consideration of US vehicles (cars, trucks, and SUVs).

Incentive-alignment. The domain was vehicles costing up to $45,000 (US). In practice, an automaker might offer an incentive-aligned lottery with the probability of winning sufficiently low that the expected cost per respondent would be reasonable. "Sweepstakes insurance" can be used to manage the risk that a vehicle would actually be awarded. (Such insurance is common in media, retail, shows, exhibits, sports promotion, and automotive promotions.)

It was not feasible in our pretest to promise respondents $45,000 toward a new vehicle, so we asked them to image they had won a lottery. Otherwise the instructions were similar to those used for mobile phones. While this pseudo-incentive-alignment might affect the predictability of the methods, it is unlikely to affect the evaluation of scalability. At this point we can only speculate about whether low odds would be sufficient for incentive alignment.

Feasibility. All 47 pretest respondents answered both the direct-elicitation and profile-evaluation tasks and all respondents were able to complete the tasks. [Recall that the 30 profiles in the profile-evaluation task were a tiny fraction of the full orthogonal design.] On a 7-point scale, with "1" = "extremely easy" and "7" = "extremely hard," respondents found the e-mail task to be relatively easy to understand (mean = 2.81, SD = 1.23) and relatively easy to write instructions to an agent (mean = 2.98, SD = 1.58). Respondents were moderately confident that the agent would select the best vehicle for them (mean = 3.21, SD = 1.40). Respondents thought they could express preferences better with the direct-elicitation task (vs. the decompositional task) as measured by a 7-point scale where a "1" meant direct elicitation was better and a "7" meant profile evaluation was better (mean = 3.09, SD = 1.98). Respondents found the direct-elicitation and profile evaluation tasks comparable on ease and realism (mean ease = 4.06, SD = 1.59; mean realism = 4.06, SD = 1.95). These results suggest that the e-mail task is practical for problems that are extremely challenging, if not infeasible, with traditional experimental designs.

Qualitatively, respondents liked the task. For example, one respondent said: "*I knew exactly what I wanted and I was told the how the purchasing phase was completed. I made sure to state my specifications but keep them in the range of the purchasers capabilities.*" While not all respondents were so positive, many were and the mean ratings reflect this. Indeed, respondents who had the most difficulty with the task were not currently in the market for a new vehicle. (The pretest was a convenience sample focused on feasibility; the automaker's profile-evaluation study will screen respondents so that they are in the market for a new vehicle.)

Based on this feasibility pretest we tentatively posit that incentive-aligned direct elicitation scales well to larger problems. This feasibility pretest also suggests that US respondents are also comfortable with the incentive-aligned direct-elicitation task.

DISCUSSION AND COMMENTARY

Together our initial comparative test and our feasibility pretest suggest a scalable method for incentive-aligned mixed-compensatory/non-compensatory direct-elicitation of consumers' decision rules for consideration-set formation. Using coded qualitative statements, combined with a simple model for a utility threshold (estimated on calibration data), incentive-aligned direct elicitation appears to predict about as well as popular incentive-aligned decompositional models. This is especially true for delayed validation that minimizes common-methods effects. It appears that unstructured direct elicitation, as implemented in our study, elicits rules that predict better than structured direct elicitation.

Discussions with market research firms with expertise in both quantitative and qualitative methods suggest that the cost of the tested direct-elicitation method is comparable to that of the decompositional methods.[9] While direct elicitation requires independent coders, such coders are

[9] Personal communication with senior officers at one such firm

often billed at lower rates than experienced quantitative analysts. Many market research firms have experienced, trained coders for qualitative data, but lack the same depth of experience for advanced statistics (although widely-available Sawtooth Software helps). If the results in this paper generalize, it appears that the choice of decomposition or direct elicitation should be made on grounds other than predictive ability or cost. Both direct elicitation and additive decomposition appear to have sufficient flexibility to model mixed non-compensatory/compensatory rules for moderately-sized experimental designs.

The biggest advantage of direct elicitation is scalability. The $4^5 2^2$ design tested in this paper was chosen to be feasible for decompositional methods. However, some applications require many more features and feature-levels. Even efficient orthogonal designs put a strain on respondents making it extremely costly (or even infeasible) to collect sufficient data. Although shrinkage in Bayesian estimation mitigates this concern, a reduced number of profiles compromises respondent-level precision for particularly large designs. On the other hand, we expect that direct elicitation should be able to handle large numbers of features and feature-levels.

One concern might be that the e-mail format could prove cumbersome if there were even more features than the automotive pretest. While this is yet to be tested, behavioral theory suggests that when faced with complex decisions involving many features, levels, or profiles, consumers often choose cognitively-simple rules and focus on a few key features (Martignon and Hoffrage 2002; Payne, Bettman, Johnson 1993; Shugan 1980). It is a reasonable hypothesis that such heuristic decision processes can be captured in an e-mail/narrative format. In direct elicitation respondents need only describe rules for the features they use to evaluate profiles. If the decision rules are simple, the number of elicited features or feature-levels will be small. On the other hand, decompositional methods tend to require a fraction of the entire experimental design.

Another concern is whether incentive compatibility can be extended from frequently-purchased products and moderately-priced durables to high-priced durables and business-to-business (B2B) products. Sweepstakes-insurance makes incentive-alignment feasible for high-priced durables. The same strategy might be tried for B2B products if the firm has already solved the agency problem so that its employees act in the best interests of the firm. We are aware of a printer manufacturer who is considering the e-mail task to identify the levels of print quality that are acceptable to commercial printers.

One final advantage of direct elicitation is the serendipitous insights that come naturally with qualitative data. By comparison decompositional methods require additional qualitative questions and the requisite coding. For example, some respondents gave reasons for their decision rules such as “rotational phones tend to break down” or “Lenovo has a younger image.” One might use incentive-aligned direct elicitation early in the product-development process to explore and finalize a product’s features and levels. This and other extensions are worth further exploration.

CAVEATS AND FUTURE DIRECTIONS

Decompositional methods are well-developed, but incentive-aligned direct elicitation for mixed non-compensatory/compensatory consideration decisions is relatively new. Our proposed methods can be improved in many ways.

Mechanism design is a key opportunity. There are proven mechanisms for willingness to pay such as the BDM procedure (Becker, DeGroot and Marschak 1964), but the intermediate decision to consider a product is a new challenge. Our incentives appear to have internal validity, motivate respondents to think hard and accurately, and are easy to understand, but they can be improved with further experimentation and experience. We would retain the prize, the dispute resolution among agents, and the agent-auditing process, but would experiment with different wordings and/or award procedures.

Training and question order is another concern. By the time the qualitative e-mails were written the respondents were encouraged to think deeply about mobile phones. This might enhance external validity, but it might also have influenced their decision rules. Such order effects are of greatest concern for the Initial Validation, but are likely mitigated for the Delayed Validation. Nonetheless, we suggest tests of order effects for the e-mail task. It is also important to optimize the amount and type of respondent training so that respondents think deeply about the product category without being induced toward artificial construction of decision rules.

Many open questions remain such as true external validity (can we predict the share of a completely new product launched to the market), scalability (to other feature-rich products and services), and really new product categories (where respondents may be more likely to use non-compensatory heuristics). Although we attempted to choose a reasonably complete set of decompositional methods for the consider-vs.-not-consider task, testing versus other decompositional methods might yield further insights. We might also improve direct elicitation with traditional or adaptive self-explication (e.g., Netzer and Srinivasan 2007). Finally, there are still challenges in finding efficient ways to summarize the managerial outputs of non-compensatory decision rules, whether they be from direct elicitation or decomposition.

REFERENCES

Akaah, Ishmael P. and Pradeep K. Korgaonkar (1983), "An Empirical Comparison of the Predictive Validity of Self-explicated, Huber-hybrid, Traditional Conjoint, and Hybrid Conjoint Models," *Journal of Marketing Research*, 20, (May), 187-197.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak (1964), "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, 9 (July), 226-232.

Bateson, John E. G., David Reibstein, and William Boulding (1987), "Conjoint Analysis Reliability and Validity: A Framework for Future Research," *Review of Marketing*, Michael Houston, Ed., pp. 451-481.

Boros, Endre, Peter L. Hammer, Toshihide Ibaraki, and Alexander Kogan (1997), "Logical Analysis of Numerical Data," *Mathematical Programming*, 79:163--190, August 1997

------, ------, ------, ------, Eddy Mayoraz, and Ilya Muchnik (2000), "An Implementation of Logical Analysis of Data," *IEEE Transactions on Knowledge and Data Engineering*, 12(2), 292-306.

Bröder, Arndt (2000), "Assessing the Empirical Validity of the `Take the Best` Heuristic as a Model of Human Probabilistic Inference," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 5, 1332-1346.

Chaloner, Kathryn and Isabella Verdinelli (1995), "Bayesian Experimental Design: A Review," *Statistical Science*, 10, 3, 273-304. (1995)

Ding, Min (2007), "An Incentive-Aligned Mechanism for Conjoint Analysis," *Journal of Marketing Research*, 54, (May), 214-223.

-----, Rajdeep Grewal, and John Liechty (2005), "Incentive-Aligned Conjoint Analysis," *Journal of Marketing Research*, 42, (February), 67–82.

Elrod, Terry, Jordan Louviere, and Krishnakumar S. Davey (1992), "An Empirical Comparison of Ratings-Based and Choice-based Conjoint Models," *Journal of Marketing Research* 29, 3, (August), 368-377.

Fishbein, Martin (1967), "Attitude and the Prediction of Behavior," in *Readings in Attitude Theory and Measurement*, Martin Fishbein, Ed. (New York, NY: John Wiley and Sons, Inc.), 477-492.

----- and Icek Ajzen (1975), *Belief, Attitude, Intention, and Behavior*, (Reading, MA: Addison-Wesley).

Frederick, Shane (2005), "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives*. 19(4). 25-42.

German, Kent (2007), "Cell phone lessons from Hong Kong," *CNET News (Crave)*, January 19, http://news.cnet.com/8301-17938_105-9679298-1.html.

Gilbride, Timothy J. and Greg M. Allenby (2004), "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," *Marketing Science*, 23(3), 391-406.

----- and ----- (2006), "Estimating Heterogeneous EBA and Economic Screening Rule Choice Models," *Marketing Science*, 25, 5, (September-October), 494-509.

Green, Paul E., (1984), "Hybrid Models for Conjoint Analysis: An Expository Review," *Journal of Marketing Research*, pp. 155-169.

----- (2004), "Thirty Years of Conjoint Analysis: Reflections and Prospects," *Conjoint Analysis, Related Modeling, and Applications: Market Research and Modeling: Progress and Prospects*, Jerry Wind and Paul Green, Eds., (Boston, MA: Kluwer Academic Publishers), 141-168.

----- and Kristiaan Helsen (1989), "Cross-Validation Assessment of Alternatives to Individual-Level Conjoint Analysis: A Case Study," *Journal of Marketing Research*, pp. 346-350.

-----, -----, and Bruce Shandler (1988), "Conjoint Internal Validity Under Alternative Profile Presentations," *Journal of Consumer Research*, 15, (December), 392-397.

-----, Abba M. Krieger, and Pradeep Bansal (1988), "Completely Unacceptable Levels in Conjoint Analysis: A Cautionary Note," *Journal of Marketing Research*, 25, (Aug), 293-300.

-----, -----, and Yoram Wind (2001), "Thirty Years of Conjoint Analysis: Reflections and Prospects," *Interfaces*, 31, 3, Part 2, (May-June), S56-S73.

Griffin, Abbie and John R. Hauser (1993), "The Voice of the Customer," *Marketing Science*, vol. 12, No. 1, (Winter), 1-27.

Hauser, John R. (1978), "Testing the Accuracy, Usefulness and Significance of Probabilistic Models: An Information Theoretic Approach," *Operations Research*, 26, 3, (May-June), 406-421

----- and Birger Wernerfelt (1990), "An Evaluation Cost Model of Consideration Sets," *Journal of Consumer Research*, 16 (March), 393-408.

-----, Olivier Toubia, Theodoros Evgeniou, Daria Silinskiai, and Rene Befurt (2008), "Cognitive Simplicity and Consideration Sets," Working Paper, MIT Sloan School of Management, Cambridge MA (October).

Hoepfl, Robert T. and George P. Huber (1970), "A Study of Self-Explicated Utility Models," *Behavioral Science*, 15, 408-414.

Hogarth, Robin M. and Natalia Karelaia (2005), "Simple Models for Multiattribute Choice with Many Alternatives: When It Does and Does Not Pay to Face Trade-offs with Binary Attributes," *Management Science*, 51, 12, (December), 1860-1872.

Huber, Joel, Dick R. Wittink, John A. Fiedler, and Richard Miller (1993), "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice," *Journal of Marketing Research*, pp. 105-114.

Hughes, Marie Adele and Dennis E. Garrett (1990), "Intercoder Reliability Estimation Approaches in Marketing: A Generalizability Theory Framework for Quantitative Data," *Journal of Marketing* Research, 27, (May), 185-195.

Jedidi, Kamel and Rajeev Kohli (2005), "Probabilistic Subset-Conjunctive Models for Heterogeneous Consumers," *Journal of Marketing Research*, 42 (4), 483-494.

Klein, Noreen M. (1986), "Assessing Unacceptable Attribute Levels in Conjoint Analysis," *Advances in Consumer Research* vol. XIV, pp. 154-158.

Kohli, Rajeev, and Kamel Jedidi (2007), "Representation and Inference of Lexicographic Preference Models and Their Variants," *Marketing Science*, 26(3), 380-399.

Kugelberg, Ellen (2004), "Information Scoring and Conjoint Analysis," Department of Industrial Economics and Management, Royal Institute of Technology, Stockholm, Sweden.

Kullback, Solomon, and Leibler, Richard A. (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79-86.

Leigh, Thomas W., David B. MacKay, and John O. Summers (1984), "Reliability and Validity of Conjoint Analysis and Self-Explicated Weights: A Comparison," *Journal of Marketing Research*, pp. 456-462.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15(2), p. 173--91.

Martignon, Laura and Ulrich Hoffrage (2002), "Fast, Frugal, and Fit: Simple Heuristics for Paired Comparisons," *Theory and Decision*, 52, 29-71.

Moore, William L. and Ekaterina Karniouchina (2006), "Screening Rules and Consumer Choice:

A Comparison of Compensatory vs. Non-Compensatory Models," Working Paper, University of Utah, Salt Lake City Utah.

----- and Richard J. Semenik (1988), "Measuring Preferences with Hybrid Conjoint Analysis: The Impact of a Different Number of Attributes in the Master Design," *Journal of Business Research*, pp. 261-274.

Netzer, Oded and V. Srinivasan (2007), "Adaptive Self-Explication of Multi-Attribute Preferences," Research Paper, Stanford Graduate School of Business, Palo Alto, CA.

Olshavsky, Richard W. and Franklin Acito (1980), "An Information Processing Probe into Conjoint Analysis," *Decision Sciences*, 11, (July), 451-470.

Park, Young-Hoon, Min Ding, Vithala R. Rao (2008) "Eliciting Preference for Complex Products: Web-Based Upgrading Method", *Journal of Marketing Research*, 45 (5), p. 562-574

Payne, John W. (1976), "Task Complexity and Contingent Processing in Decision Making: An Information Search," *Organizational Behavior and Human Performance*, 16, 366-387.

-----, James R. Bettman and Eric J. Johnson (1988), "Adaptive Strategy Selection in Decision Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3), 534-552.

-----, ----- and ----- (1993), *The Adaptive Decision Maker*, (Cambridge UK: Cambridge University Press).

Perreault, William D., Jr. and Laurence E. Leigh (1989), "Reliability of Nominal Data Based on Qualitative Judgments," *Journal of Marketing Research*, 26, (May), 135-148.

Prelec, Dražen (2004), "A Bayesian Truth Serum for Subjective Data," *Science*, 306, (October 15), 462-466.

Roberts, John H., and James M. Lattin (1991)," Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 28 (November), 42940.

Rossi, Peter E., Greg M. Allenby (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22(3), p. 304-328.

Sawtooth Software, Inc. (1996), "ACA System: Adaptive Conjoint Analysis," *ACA Manual*, (Sequim, WA: Sawtooth Software, Inc.)

----- (2004), "The CBC Hierarchical Bayes Technical Paper," (Sequim, WA: Sawtooth Software, Inc.)

Shugan, Steven (1980), "The Cost of Thinking," *Journal of Consumer Research*, 27(2), 99-111.

Silinskaia, Daria, John R. Hauser, and Glen L. Urban (2009), "Adaptive Profile Evaluation to Identify Heuristic Decision Rules in Large and Challenging Experimental Designs," MIT Sloan Working Paper, Cambridge, MA.

Smith, Vernon L. (1976), "Experimental Economics: Induced Value Theory," *American Economic Review*, 66 (May), 274-79.

Srinivasan, V. (1988), "A Conjunctive-Compensatory Approach to The Self-Explication of Multiattributed Preferences," *Decision Sciences*, pp. 295-305.

----- and Chan Su Park (1997), "Surprising Robustness of the Self-Explicated Approach to Customer Preference Structure Measurement," *Journal of Marketing Research*, 34, (May), 286-291.

----- and Gordon A. Wyner (1988), "Casemap: Computer-Assisted Self-Explication of Multiattributed Preferences," in W. Henry, M. Menasco, and K. Takada, Eds, *Handbook on New Product Development and Testing*, (Lexington, MA: D.C.Heath), 91-112.

Swait, Joffre and Tülin Erdem (2007), "Brand Effects on Choice and Choice Set Formation Under Uncertainty," *Marketing Science* 26, 5, (September-October), 679-697.

Toubia, Olivier (2006), "Idea Generation, Creativity, and Incentives," *Marketing Science*, 25, 5, (September-October), 411-425.

-----, John R. Hauser and Rosanna Garcia (2007), "Probabilistic Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis: Theory and Application," *Marketing Science*, 26, 5, (September-October), 596-610.

-----, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, 22(3), 273-303.

Wilkie, William L. and Edgar A. Pessemier (1973), "Issues in Marketing's Use of Multi-attribute Attitude Models," *Journal of Marketing Research*, 10, (November), 428-441.

Wind, Jerry, Paul E. Green, Douglas Shifflet, and Marsha Scarbrough (1989), "Courtyard by Marriott: Designing a Hotel Facility with Consumer-Based Marketing Models," *Interfaces*, pp. 25-47.

Wright, Peter (1973), "The Cognitive Processes Mediating Acceptance of Advertising," *Journal of Marketing Research*, 10, (February), 53-62.

Yee, Michael, Ely Dahan, John R. Hauser and James Orlin (2007) "Greedoid-Based Noncompensatory Inference," *Marketing Science*, 26, 4, (July-August), 532-549.

TABLE 1. PREDICTIVE ABILITY

Initial Validation	Hit Rate	Relative Improvement	Consideration-Set Size	K-L Percentage
Decompositional Methods				
HB Logit, Additive Utility	**81.3%***	**55.1%***	8.6	**26.6%***
HB Logit, *q*-Compensatory	79.1%	49.7%	7.8	20.7%
Greedoid Dynamic Program[1]	78.1%	47.3%	9.0	**25.8%***
Logical Analysis of Data[2]	78.7%	48.8%	8.1	24.6%
Direct-Elicitation Methods				
Raw Conjunctive	63.4%	11.9%	17.7	24.8%
Modified Conjunctive	76.2%	42.8%	6.1	22.2%
Mixed, Match Cutoff	79.5%	50.7%	8.2	**29.0%***
Mixed, Logit-based Cutoff	77.2%	45.1%	8.6	**28.3%***

Delayed Validation	Hit Rate	Relative Improvement	Consideration-Set Size	K-L Percentage
Decompositional Methods				
HB Logit, Additive Utility	**79.5%***	**50.9%***	8.5	**27.0%***
HB Logit, *q*-Compensatory	76.9%	44.5%	8.8	21.0%
Greedoid Dynamic Program[1]	76.3%	43.2%	9.0	**26.1%***
Logical Analysis of Data[2]	77.2%	45.4%	8.1	**25.6%***
Direct-Elicitation Methods				
Raw Conjunctive	62.3%	9.6%	17.8	24.1%
Modified Conjunctive	76.5%	43.6%	6.2	23.5%
Mixed, Match Cutoff	**78.7%***	**49.0%***	8.2	**28.7%***
Mixed, Logit-based Cutoff	77.5%	46.0%	8.7	**28.0%***

[1] Estimates a lexicographic model. [2] Estimates disjunctive, conjunctive, subset conjunctive, and/or disjunctions of conjunctions models. * Best in column or not significantly different than best at the 0.05 level.

TABLE 2: RULES AND PARTWORTHS BY FEATURE LEVEL

Feature	Level	Direct Elicitation Percent Elimination	Direct Elicitation Percent Compensatory	Decomposition HB Mean Partworths[1]	HB Partworth Heterogeneity (Std Dev)[2]
Brand	Motorola	12.6%	14.7%	—	—
	Lenovo	15.4%	13.3%	-0.233	0.500
	Nokia	1.4%	60.1%	1.135	0.354
	Sony-E	3.5%	48.3%	0.833	0.406
Color	Black	2.8%	53.8%	—	—
	Blue	8.4%	24.9%	-0.423	0.393
	Silver	0.7%	46.2%	0.068	0.751
	Pink	29.4%	21.7%	-2.073	2.354
Screen Size	Small	16.8%	0.0%	—	—
	Large	0.0%	79.0%	2.380	1.618
Thickness	Slim	0.0%	51.0%	—	—
	Normal	7.0%	4.9%	-0.629	0.413
Resolution	0.5 Mp	31.5%	14.0%	—	—
	1.0 Mp	23.8%	25.2%	1.021	0.422
	2.0 Mp	3.5%	69.2%	3.348	1.738
	3.0 Mp	0.0%	81.1%	3.731	2.122
Style	Bar	5.6%	43.4%	—	—
	Flip	8.4%	34.3%	-0.127	0.411
	Slide	4.9%	42.0%	0.076	0.391
	Rotational	16.8%	28.7%	-0.581	0.960
Price		18.9%	2.8%	—	—
Base Price	$HK1080	—	—	—	—
	$HK1280	—	—	-0.095	0.136
	$HK1480	—	—	-0.031	0.401
	$HK1680	—	—	-0.167	0.307

[1] Posterior mean of the partworths from the decompositional “HB Logit, Additive Utility” model.
[2] Posterior partworth standard deviation (across respondents) from the “HB Logit, Additive Utility” model.

SUPPLEMENTAL APPENDIX (Available from the authors)

ANALYSIS OF CHOICE WITHIN THE CONSIDERATION SET

In the text we focus on respondents' decisions with respect to which mobile phone they would consider. We chose this focus for ease of exposition, 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 becoming an important managerial problem.

Our study also asked respondents to rank profiles within their consideration sets. Three of the four decompositional methods rank order the profiles and the mixed-rule direct-elicitation method weakly orders the profiles. From these predicted ranks we compute the rank correlation with the observed ranks in the validation data. Table A1 summarizes these results.

The data are consistent with Table 1 in the text in that there is no statistical difference between the decompositional additive logit method and the mixed-rule direct-elicitation method, this time 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, rather than choice – an hypothesis worth further testing.

TABLE A1. RANK CORRELATIONS FOR CHOICE WITHIN CONSIDERATION SET

	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		
Mixed Decision Rules	**0.412***	**0.375***

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

SUPPLEMENTAL APPENDIX (Available from the authors)

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.

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; 2000). 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.