MIN DING, JOHN R. HAUSER, SONGTING DONG, DARIA DZYABURA, ZHILIN YANG, CHENTING SU, and STEVEN P. GASKIN*

The authors investigate the feasibility of unstructured direct elicitation (UDE) of decision rules consumers use to form consideration sets. They incorporate incentives into the tested formats that prompt respondents to state noncompensatory, compensatory, or mixed rules for agents who will select a product for the respondents. In a mobile phone study, two validation tasks prompt respondents to indicate which of 32 mobile phones they would consider from a fractional design of features and levels. The authors find that UDE predicts consideration sets better, across both profiles and respondents, than a structured direct-elicitation method. It predicts comparably to established incentive-aligned compensatory, noncompensatory, and mixed decompositional methods. In a more complex automotive study, noncompensatory decomposition is not feasible and additive-utility decomposition is strained, but UDE scales well. The authors align incentives for all methods using prize indemnity insurance to award a chance at $40,000 for an automobile plus cash. They conclude that UDE predicts consideration sets better than either an additive decomposition or an established structured direct-elicitation method (CASEMAP).

*Keywords*: decision rules, consideration sets, direct elicitation, incentive alignment, product development

# Unstructured Direct Elicitation of Decision Rules

We explore direct elicitation of decision rules that have the potential to scale to domains that challenge decompositional approaches. These incentive-aligned approaches encourage consumers to self-state both compensatory and noncompensatory rules and recognize that consumers often use a consider-then-choose process, especially in complex product categories. (Our primary focus is on the consideration decision.) We study an unstructured mechanism in which a consumer composes an e-mail that “teaches” an agent how to make decisions for the consumer. Following current best practices, we align incentives for both the consumer and the agent so that the consumer is motivated to think hard and provide accurate answers.

Two complementary experiments compare unstructured direct elicitation (UDE) with decompositional and self-explication approaches that have proved successful in other empirical comparisons. The first experiment is in a category (mobile phones, $4^5 \times 2^2$ design) in which most decompositional approaches are feasible. The teach-an-agent task predicts consideration as well as a standard hierarchical Bayes (HB) additive logit model and establishes noncompensatory decompositional decision models but better than a pure compensatory decompositional model. We also conclude that an unstructured teach-an-agent task does better than one in which we force structure. The second experiment is in a category (automobiles, $20 \times 7 \times 5^2 \times 4 \times 3^4 \times 2^2$ design) in which noncompensatory decomposition is not feasible and standard decomposition methods are challenged. We find that UDE scales well to this application and predicts

*Min Ding is Smeal Research Fellow in Marketing and Professor of Marketing, Smeal College of Business, Pennsylvania State University (e-mail: minding@psu.edu). John R. Hauser is Kirin Professor of Marketing (e-mail: hauser@mit.edu), and Daria Dzyabura is a doctoral student (e-mail: dariasil@mit.edu), MIT Sloan School of Management, Massachusetts Institute of Technology. Songting Dong is Lecturer in Marketing, Research School of Business, Australian National University (e-mail: dongst@gmail.com). Zhilin Yang (e-mail: mkzyang@cityu.edu.hk) and Chenting Su (e-mail: mkctsu@cityu.edu.hk) are Associate Professors of Marketing, City University of Hong Kong. Steven P. Gaskin is a principal, Applied Marketing Science, Inc. (e-mail: sgaskin@ams-inc.com). The authors acknowledge financial support from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project 9041182, CityU 1454/06H; Project 9041414, CityU 148608), and the Smeal Small Research Grant from Pennsylvania State University, as well as constructive comments from Michael Braun, Ely Dahan, John Liechty, Young-Hoon Park, Vithala Rao, and Kayande Ujwal and from participants in a seminar given at University of Houston and a presentation given at the 2009 Marketing Science Conference in Ann Arbor, Michigan. Carl Mela served as associate editor for this article.

ISSN: 0022-2437 (print), 1547-7193 (electronic)

better than HB logit analysis. It also predicts better than an established structured direct-elicitation (SDE) approach (CASEMAP; e.g., Srinivasan 1988). To maintain consistency, we aligned incentives for all tested approaches, even automobiles, for which respondents had a reasonable chance of getting a task-defined $40,000 automobile (plus cash if the automobile was priced less than $40,000).

Our research goals are proof of concept and initial test. Our goal is to demonstrate that a UDE method can be designed to be incentive aligned and that, in some circumstances, UDE will predict consideration as well as or better than most commonly used decompositional and compositional methods. We choose benchmarks that use a variety of methods and have done well in previous comparative testing.

## *MOTIVATION*

This research is motivated by five advances in behavioral theory and managerial practice. First, applications such as automobiles and high-technology gadgets have become rich in features, requiring large numbers of profiles for even orthogonal experimental designs. For example, Dzyabura and Hauser (2010) describe a study a U.S. automaker used that would have required a minimal orthogonal design of 13,320 profiles. Our goal is to find methods that scale well to such complex applications.

Second, in Web-based purchasing, catalogs, and superstores, consumers often select from among 20–100+ products. Behavioral research suggests that, when faced with so many alternatives, consumers use a two-stage consider-then-choose process rather than a one-stage compensatory evaluation (e.g., Hauser and Wernerfelt 1990; Payne 1976; Roberts and Lattin 1991; Swait and Erdem 2007). Consumers often consider only a small fraction (<10%) of the brands available. Our goal is to find methods that capture the consider-then-choose decision process. (In this study, we focus primarily on the consideration stage, relegating the choice stage to exploratory results in the Web Appendix at http://www.marketingpower.com/jmrfeb11 and further research.)

Third, behavioral research and decompositional methods suggest that, particularly when faced with many feature-rich products, 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). Our goal is to find methods that measure both compensatory and noncompensatory decision rules.

Fourth, recent research suggests that incentive alignment, through natural tasks that consumers do in their daily lives with actual consequences, leads to greater respondent involvement, less boredom, and higher data quality (Ding 2007; Ding, Grewal, and Liechty 2005; Ding, Park, and Bradlow 2009; Kugelberg 2004; Park, Ding, and Rao 2008; Prelec 2004; Smith 1976; Toubia, Hauser, and Garcia 2007; Toubia et al. 2003). In theory, incentive alignment gives consumers sufficient motivation to describe their decision rules accurately. For a fair comparison with established methods, we accept incentive alignment as state of the art and induce incentives for the proposed and established methods. We leave comparisons when incentives are not aligned to further research.

Fifth, 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). Although the labor cost for such coding is linear in the number of respondents, voice-of-the-customer experience suggests that for typical sample sizes, the costs of lower-wage coders roughly balance the fixed cost of the higher-wage analysts who are necessary for the analysis of decompositional data. (This is not surprising. Market forces have led to efficiencies so that both voice of the customer and decomposition can compete in the market.) Coding costs increase linearly with sample size but not with the complexity of the product category because, empirically, consumers often strive for simplicity in their heuristic decision rules (Gigerenzer and Goldstein 1996; Payne, Bettman, and Johnson 1988, 1993).

## *PREVIOUS LITERATURE*

Direct elicitation (sometimes called self-explication or composition) has been used to measure consumer preferences and/or attitudes for more than 40 years either alone or in combination with decompositional methods (Fishbein and Ajzen 1975; Green 1984; Hoepfl and Huber 1970; Sawtooth Software 1996; 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; Hauser and Wisniewski 1982; Huber et al. 1993; Leigh, MacKay, and Summers 1984; Moore and Semenik 1988; Srinivasan and Park 1997). Attempts at the SDE of noncompensatory rules have met with less success, partly because respondents often choose profiles with levels they say are "unacceptable" (Green, Krieger, and Bansal 1988; Klein 1986; Sawtooth Software 1996; Srinivasan and Wyner 1988).

Researchers have proposed decompositional methods for conjunctive, disjunctive, subset conjunctive, lexicographic, and disjunctions of conjunctions decision rules (Gilbride and Allenby 2004, 2006; Hauser et al. 2011; Jedidi and Kohli 2005; Kohli and Jedidi 2007; Moore and Karniouchina 2006; Yee et al. 2007).[1] The results to date suggest that noncompensatory methods predict comparably to, but sometimes less well than, compensatory methods in product categories with which respondents are familiar (e.g., batteries, computers). Noncompensatory methods are slightly better in unfamiliar categories (e.g., smart phones, global positioning systems). Research suggests that approximately one-half to two-thirds of the respondents are fit better with noncompensatory than compensatory methods and that the percentage is higher when respondents are asked to evaluate

[1] 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.

more profiles. The vast majority of identified heuristics are conjunctive rules (Hauser et al. 2011). The results are comparable whether the decision is consideration, consider-then-choose, or choice. We are unaware of any comparisons with noncompensatory direct-elicitation methods.

## *THE MOBILE PHONE STUDY*

It is easier to describe the direct-elicitation and decompositional tasks through examples; therefore, we begin with a brief description of the product category used in the first study. 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 (German 2007). Using local informants, observation of mobile phone stores, and discussions with potential respondents, we selected a set of features and feature levels that represent the choices Hong Kong respondents face. Pretests indicated that the following feature levels were face valid:

- Brand: Motorola, Lenovo, Nokia, or Sony-Ericsson
- Color: black, blue, silver, or pink
- Screen size: small (1.8 in.) or large (3.0 in.)
- Thickness: slim (9 mm) or normal (17 mm)
- Camera resolution: .5 MP, 1.0 MP, 2.0 MP, or 3.0 MP
- Style: bar, flip, slide, or rotational
- Base price level: HK$1080, HK$1280, HK$1480, or HK$1680 (US$1 ≈ HK$8)

This $4^5 \times 2^2$ design is typical of compensatory decompositional analysis and at the upper limit of noncompensatory decompositional methods requiring computations that are exponential in the number of feature levels.

### *Direct-Elicitation Tasks for the Mobile Phone Study*

We developed two direct-elicitation tasks in the mobile phone study. First, an SDE task prompted respondents to provide rules for a friend who would act as their agent in considering and/or purchasing a product for them. Respondents were asked to state instructions unambiguously and to use as many instructions as necessary. The task format had open boxes for five rules, though respondents were not required to state five rules and they could add rules if desired. Second, a UDE task prompted respondents to state their instructions to the agent in the form of an e-mail to a friend. Other than a requirement to begin the e-mail with "Dear friend," respondents could use any format to describe their decision rules.

Two independent judges, who were blind to any hypotheses, coded each direct-elicitation task independently. Then, the two judges met to reconcile differences. Such coding is common in market research for both commercial use and litigation (e.g., Hughes and Garrett 1990; Perreault and Leigh 1989; Wright 1973). (The coding guide, the transcripts, and all coded responses are available on request.)

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

To illustrate the coding, we provide example statements from respondents' e-mails (retaining original language and grammar):

> [Mostly noncompensatory] 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, .5 MP, Motorola, Lenovo, blue, and pink; 1 for 3.0 MP.]

> [Mixed noncompensatory/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 [be] 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 1.0 MP. 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 .5 MP, pink, and small screen; 1 for slide and rotational; and 4 for Lenovo. Note that because our coding is conservative, for this respondent, neither the subjective statements of relative importance of features nor the target price was 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: 2 MP 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 [*sic*] 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.]

### *Decompositional Task*

We based the decompositional benchmark models on a three-panel format that Hauser and colleagues (2011) develop. The left panel shows icons representing the 32 mobile phones. Respondents chose profiles from an orthogonal fractional factorial of the $4^5 \times 2^2$ design. When the respondent clicked on an icon, the mobile phone appeared in the center panel. (Pictures and text described the features.) The respondent indicated whether 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 or her consideration set. The data to estimate the decompositional models are 0 versus 1 indicators of whether each profile is included in the consideration set.

To make the respondent's task realistic and avoid dominated profiles (e.g., Elrod, Louviere, and Davey 1992; Johnson, Meyer, and Ghose 1989), we set the price levels for each profile as 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 added $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 do they inhibit the noncompensatory use of price (Green, Helsen, and Shandler 1988; Hauser et al. 2011; Toubia et al. 2003; Toubia, Hauser, and Simester 2004).

### *Benchmark Compensatory, Noncompensatory, and Mixed Models*

We chose commonly used compensatory and noncompensatory decompositional methods as benchmarks. Our first benchmark is the standard HB logit model applied to consideration sets using the 32 consider-versus-not-consider observations per respondent (Hauser et al. 2011; Lenk et al. 1996; Rossi and Allenby 2003; Sawtooth Software 2004; Swait and Erdem 2007). The specification is an additive partworth model. Many researchers have argued that compensatory, lexicographic, 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).[2] Following Bröder (2000) and Yee and colleagues (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 the 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 noncompensatory decision rules.

There are a variety of noncompensatory decompositional models and estimation methods to use as benchmarks. We selected two that have done well in previous research: the greedoid dynamic program, which estimates a lexicographic consideration set model (Yee et al. 2007), and logical analysis of data, which estimates disjunctions of conjunctive rules (Boros et al. 1997). Disjunctions of conjunctive rules are generalizations of disjunctive, conjunctive, subset conjunctive, and, in the case of consideration data, lexicographic rules. Logical analysis of data has matched or outperformed other noncompensatory decompositional methods, including HB specifications of conjunctive, disjunctive, and subset conjunctive models, in at least one study (Gilbride and Allenby 2004, 2006; Hauser et al. 2011; Jedidi and Kohli 2005). We hope that together the two methods provide reasonable initial benchmarks to represent a broader set of noncompensatory decompositional methods. (For a summary of the benchmark methods, see the Web appendix at http://www.marketingpower.com/jmrfeb11.)

### *Participants and Study Design*

The participants were students at a major university in Hong Kong who were screened to be at least 18 years of age and interested in purchasing a mobile phone. After a pretest, in which 56 respondents indicated that the questions were clear and the task not onerous, we invited participants to come to a computer laboratory on campus to complete the Web-based survey. They also completed a delayed validation task on any Internet-connected computer three weeks later. Those who completed both tasks received $HK100 and were eligible to receive an incentive-aligned prize (as we describe subsequently). 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 rather than the choice task because (1) there is growing managerial and scientific interest in consideration decisions, (2) direct elicitation of consideration rules is relatively novel in the literature, and (3) the consideration task was more likely to provide a test of compensatory, noncompensatory, and mixed decision rules. Initial tests (available in the Web appendix at http://www.marketingpower.com/jmrfeb11) suggest that the predictive ability of the choice task for mobile phones (rank order within the consideration set) mimics the basic results obtained for the consideration task.

To obtain greater statistical power, we used a within-subjects design in which participants completed both direct-elicitation and decompositional tasks. We use two validation tasks: One task occurs toward the end of the Web-based survey after a memory-cleansing task, and the other task was delayed by three weeks. The validation tasks use an interface identical to the decompositional task so that common methods effects likely favor decompositional rather than direct elicitation. 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:

1. Initial screens ensured privacy and described the basic study.
2. The next screens introduced mobile phone features one at a time through text and pictures.
3. Incentives were described for both the decompositional and the direct-elicitation tasks.
4. The order of the following two tasks was randomized: (a) Respondents indicated which of 32 mobile phones they would consider (Calibration Task) and then ranked the considered profiles afterward, and (b) respondents described decision rules to be used by an agent to select a mobile phone for the respondent (SDE task).
5. Brainteaser distraction questions cleared short-term memory (Frederick 2005).

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

6. Respondents saw a new orthogonal set of 32 mobile phones (same for all respondents), indicated those they would consider (Initial Validation Task), and then ranked the considered profiles afterward.
7. Respondents wrote an e-mail as an alternative way to instruct an agent to select a mobile phone (e-mail-based UDE task).
8. Short questions measured respondents' comprehension of the incentives and tasks.
9. Three weeks later, respondents saw a third orthogonal set of 32 mobile phones (same for all respondents), indicated those they would consider (Delayed Validation Task), and then ranked the considered profiles afterward.

This design focuses on methods comparison. At a minimum, we believe that 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 an actual 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 that the incentives also enhance external validity. At a minimum, pretest comments and postsurvey debriefs suggest that respondents believed they would behave in the market as they did in the survey. (Moreover, respondents who received mobile phones as part of the incentive were satisfied with the phones agents chose for them.)

A second concern is that either the decompositional estimation task or the direct-elicitation task trains respondents, perhaps affecting how they 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, internal consistency would enhance both decompositional and direct-elicitation methods, perhaps favoring decomposition more because we use the same type of task for validation.

A third concern is an order effect for the UDE task (the e-mail task), which occurs after the initial validation task. Potential order effects might be mitigated for the delayed validation task, but this caveat remains for the mobile phone study. Our second study randomizes the order of the tasks and provides insight on the value of training effects (order effects).

### *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 that they would get every mobile phone considered, the best response would be a large consideration set. If we told respondents that they would receive their most-preferred mobile phone, the best response would be a consideration set of exactly one mobile phone. Instead, on the basis of pretests, we chose the following two-stage mechanism. Because this mechanism is a heuristic, we call it "incentive aligned" rather than the more formal term "incentive compatible." Our goals with incentive alignment are to ensure that the respondents believe (1) it is in their best interests to think hard and tell the truth; (2) it is, as much as feasible, in their best interests to do so; and (3) there is no way, that is obvious to the respondents, they can improve their welfare by "cheating."

Specifically, we told respondents that they had a 1 in 30 chance of receiving a mobile phone plus cash representing the difference between the price of the phone and HK$2500.[3] Because we wanted both the direct-elicitation and the decompositional tasks to be incentive aligned, we told respondents that one of the tasks would be selected by a coin flip 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, we told respondents that we would first randomly select one of the three tasks (two in the main study and one 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 more than one phone matched their consideration set, the rank data would distinguish the phones. The unknown random subset is important here. This design reflects a real-life scenario in which a consumer constructs his or her consideration set knowing that random events, such as decreased product availability, can occur before purchase. 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 focus) and choice within the consideration set (see the Web appendix at http://www.marketingpower.com/jmrfeb11).

For the direct-elicitation tasks, we told respondents 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, we told respondents that the agents would be audited and not paid unless they followed the respondents' instructions accurately (e.g., Toubia 2006).

At the conclusion of the study, we selected five respondents randomly. Each received a specific mobile phone (and cash) according to the mechanism described previously. 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." There were no significant differences between the two tasks. On average, respondents found the tasks and incentive alignment easy to understand. Quali-

[3] The prize of HK$2500, approximately US$300+, might induce a wealth-endowment effect, making the respondent more likely to choose more features. Although the wealth-endowment effect 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 and colleagues (2003) endowed all respondents with $100. They report good external validity when forecasting market shares after the product was launched to the market.

tative statements also suggested that respondents believed that their answers should be truthful and reflect their true consideration decisions. (For details, see the Web appendix at http://www.marketingpower.com/jmrfeb11.)

We compare direct elicitation and decomposition when both are incentive aligned and leave for further research comparative tests when incentives are not aligned. Interactions between task and incentives would be scientifically interesting. For example, Kramer (2007) suggests that respondents trust researchers more when the task is more transparent.

## *MOBILE PHONE STUDY 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 (9.4) and Delayed Validation (9.3) tasks. All are statistically equivalent, consistent with the hypothesis that respondents thought carefully about the tasks.

The judges' classifications of directly elicited statements indicated that more than three-fourths of the respondents (78.3%) asked their agents to use a mixture of compensatory and noncompensatory rules for consideration and/or choice. Most of the remainder were compensatory (21.0%), and only one was purely noncompensatory (.7%).

### *Predictive Performance 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 because respondents consider a relatively small set of profiles. With average consideration sets of approximately 9.3 of 32 (29.1%), a null model that predicts that no mobile phones will be considered will achieve a hit rate of 70.9%. Furthermore, hit rates merge false positives and false negatives. To distinguish results from an all-reject null model, we could examine whether we predict the size of the consideration set correctly. However, an (alternative) null model of random prediction (proportional to consideration set size) gets the consideration set size correct.

Instead, we use a version of the Kullback–Leibler (KL) divergence (also known as relative entropy), which measures the expected divergence in Shannon's information measure between the validation data and a model's predictions (Chaloner and Verdinelli 1995; Kullback and Leibler 1951). This version of KL divergence 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. It discriminates among models even when the hit rates might otherwise be equal. Because it is difficult to interpret the units (bits) of KL divergence, we rescale the measure relative to the KL divergence between the validation data and a random model. (On this relative measure, larger is better. A random model has a relative KL of 0%, and perfect prediction has a relative KL of 100%.) This rescaling does not affect either the relative comparisons or the results of the statistical tests in this study.

We derive a KL formula that is comparable for both 0 versus 1 and probabilistic predictions (see the Web Appendix at http://www.marketingpower.com/jmrfeb11). The formula, which we apply to each respondent's data, aggregates to false positives, true positives, false negatives, and true negatives. Specifically, 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. The KL formula is given by the following:

$$\text{(1)} \quad \begin{aligned} KL = {} & \hat{C}_v \log_2 \hat{C}_v + (V - \hat{C}_v)\log_2(V - \hat{C}_v) \\ & - (\hat{C}_v - F_p)\log_2(\hat{C}_v - F_p) - F_n \log_2 F_n - F_p \log_2 F_p \\ & - (V - \hat{C}_v - F_n)\log_2(V - \hat{C}_v - Fn). \end{aligned}$$

The KL divergence evaluates cross-profile predictions. Elrod (2001) argues that it is also important to make comparisons between respondents and proposes a likelihood-based analysis for probabilistic predictions. For a measure that is comparable for discrete and probabilistic predictions, we bifurcate the sample and report the root mean square error (RMSE) between predictions from each half to the observed validation consideration shares in the other half (smaller is better). The RMSE between the observed consideration shares in the two half samples (Initial Validation: .083; Delayed Validation: .068,) provides a lower bound on what might be obtained with a predictive model. Because RMSE is an aggregate measure and the models are not nested, we cannot compute statistical significance for this aggregate measure.

*Predicting with directly elicited rules*. To make predictions, we use both the explicit elimination rules and the compensatory statements that weakly order noneliminated profiles.[4] The order is weak because the qualitative statements may not distinguish trade-offs among features or levels within features (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 the benefits of a larger consideration set with the cognitive costs. We do this in two ways. First, 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.

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

Because our goal is proof of concept, we believe we are justified in using consideration set sizes from the decompositional data to calibrate the binary logit model. For UDE-only applications, we suggest that the threshold model be calibrated with a pretest decompositional task or that a more efficient task be developed to elicit consideration set sizes. Until such pretest tasks are developed and tested, the

[4] Models based on both noncompensatory and compensatory statements outperformed models based on the elimination rules only and did so on all measures. Details are available on request.

reduced-data advantage of UDE for modest experimental designs is somewhat mitigated. We return to this issue in our second study, in which respondents cannot evaluate all 25,600 orthogonal profiles, severely straining additive decomposition. (Noncompensatory decomposition is not feasible in that complex of a design.)

*Comparisons*. Table 1 summarizes the predictive tests. The UDE task does significantly better than the SDE task on all comparisons. It seems that the e-mail task is more natural, making it easier for respondents to articulate their decision rules.

The best decompositional method is the HB logit with additive utility. It does substantially better on RMSE than the other decompositional methods and better, though not significantly so, on KL. From the qualitative observation that most directly elicited statements contain both compensatory and noncompensatory instructions, it is not surprising that the mixed (additive) decomposition model does well.

When we compare decompositional methods with direct-elicitation methods, we find that the direct-elicitation models are best on KL, though not significantly so. The two best models on RMSE seem to be the mixed (additive) decompositional model and the estimated-cutoff UDE model, with the former doing slightly better on the initial validation and the latter doing slightly better on the delayed validation. Notably, the RMSE for these models is only slightly larger than the lower bound on RMSE. Decomposition and direct elicitation are statistically (KL) and substantially (RMSE) better than the null models. Respondents seem to use at least some noncompensatory decision rules. Only the q-compensatory model is significantly worse on KL.

The results in Table 1 lead us to tentatively conclude the following for consideration decisions:

- The UDE task provides better data than the SDE task;
- UDE predicts comparably to the best decompositional method (of those tested) on cross-profile and cross-sample validation;
- In cross-sample validation, the best UDE and the best decompositional models come close to the lower bound, as indicated by split-half sample agreement; and
- The mobile phone respondents mix elimination and compensatory decision rules.

These are important findings, especially if UDE scales better than decomposition for applications with large numbers of features and feature levels. Because incentive-aligned direct-elicitation methods for consideration set decisions are comparatively new relative to incentive-aligned decomposition, we expect them to improve with further application.

*Other comments*. The decompositional noncompensatory models are comparable to the additive model, superior to the q-compensatory model, and superior to analyses that use only the directly elicited elimination statements. (Table 1 does not show the latter. These achieve KL percentages of 14.9% and 14.5% for the initial and delayed validations, respectively.) This predictive performance is consistent with Yee and colleagues' (2007) results.

## *ILLUSTRATIVE MANAGERIAL OUTPUTS: MOBILE PHONES*

Researchers have developed the managerial presentation of decompositional additive partworths through decades of application. The last two columns of Table 2 provide a commonly used format: the posterior means and standard deviations (across respondents). For example, on average, the pink color has a large negative partworth, but not all respondents agree: Heterogeneity among respondents is large. The HB logit, additive utility model suggests that respondents vary considerably in their preferences for most mobile phone features.

Academics and practitioners are still evolving the best way to summarize noncompensatory decision rules for managerial insight. Table 2 provides one potential summary. The third column reports the percentage of respondents

Table 1
PREDICTIVE ABILITY MOBILE PHONE STUDY

| | *Initial Validation* | | *Delayed Validation* | |
|---|---|---|---|---|
| | *Relative KL Divergence*[a] *(%)* | *Cross-Validation RMSE*[b] | *Relative KL Divergence (%)* | *Cross-Validation RMSE* |
| *Decompositional Methods* | | | | |
| HB logit, additive utility | 25.3* | .088 | 23.7* | .089 |
| HB logit, q-compensatory | 19.3 | .144 | 17.6 | .127 |
| Greedoid dynamic program[c] | 24.5* | .136 | 23.0* | .118 |
| Logical analysis of data[d] | 23.2* | .140 | 22.4* | .133 |
| *SDE* | | | | |
| Match cutoff | 19.5 | .125 | 19.7 | .110 |
| Estimated cutoff | 20.0 | .118 | 19.2 | .110 |
| *UDE* | | | | |
| Match cutoff | 27.6* | .103 | 25.4* | .100 |
| Estimated cutoff | 27.1* | .094 | 24.8* | .088 |
| *Null Models* | | | | |
| Reject all | .0 | .370 | .0 | .364 |
| Random proportional to consideration share in calibration data | .0 | .228 | .0 | .219 |
| Split-half predicted versus observed profile share cross-validation | — | .083 | — | .068 |

*Best in column or not significantly different from best in column at the .05 level.

[a] Rescaled, such that larger numbers are better.

[b] Bifold cross-validation compares predictions of profile shares from each half of the sample with profile shares in the remaining half. Smaller numbers are better.

[c] The greedoid dynamic program estimates a lexicographic model.

[d] The logical analysis of data estimates disjunctive, conjunctive, subset conjunctive, and/or disjunctions of conjunctions models.

Table 2

RULES AND PARTWORTHS BY FEATURE LEVEL: MOBILE PHONES

| *Feature/ Level* | *Direct Elicitation Percent Elimination (%)* | *Direct Elicitation Percent Compensatory (%)* | *Decomposition HB Mean Partworths*[a] | *HB Partworth Heterogeneity (SD)*[b] |
|---|---|---|---|---|
| *Brand* | | | | |
| Motorola | 12.6 | 14.7 | — | — |
| Lenovo | 15.4 | 13.3 | –.233 | .500 |
| Nokia | 1.4 | 60.1 | 1.135 | .354 |
| Sony-Ericsson | 3.5 | 48.3 | .833 | .406 |
| *Color* | | | | |
| Black | 2.8 | 53.8 | — | — |
| Blue | 8.4 | 24.9 | –.423 | .393 |
| Silver | .7 | 46.2 | .068 | .751 |
| Pink | 29.4 | 21.7 | –2.073 | 2.354 |
| *Screen Size* | | | | |
| Small | 16.8 | .0 | — | — |
| Large | .0 | 79.0 | 2.380 | 1.618 |
| *Thickness* | | | | |
| Slim | .0 | 51.0 | — | — |
| Normal | 7.0 | 4.9 | –.629 | .413 |
| *Resolution* | | | | |
| .5 MP | 31.5 | 14.0 | — | — |
| 1.0 MP | 23.8 | 25.2 | 1.021 | .422 |
| 2.0 MP | 3.5 | 69.2 | 3.348 | 1.738 |
| 3.0 MP | .0 | 81.1 | 3.731 | 2.122 |
| *Style* | | | | |
| Bar | 5.6 | 43.4 | — | — |
| Flip | 8.4 | 34.3 | –.127 | .411 |
| Slide | 4.9 | 42.0 | .076 | .391 |
| Rotational | 16.8 | 28.7 | –.581 | .960 |
| *Price* | 18.9 | 2.8 | — | — |
| *Base Price* | | | | |
| HK$1080 | — | — | — | — |
| HK$1280 | — | — | –.095 | .136 |
| HK$1480 | — | — | –.031 | .401 |
| HK$1680 | — | — | –.167 | .307 |

[a] Posterior mean of the partworths from the decompositional HB logit, additive utility model.

[b] Posterior partworth standard deviation (across respondents) from the HB logit, additive utility model.

whose directly elicited decision rules include a feature level as an elimination criterion. For example, 12.6% of the respondents mentioned that they would eliminate any Motorola mobile phone, whereas only 1.4% would eliminate any Nokia mobile phone. The highest noncompensatory feature levels are low camera resolutions (31.5%) and the pink color (29.4%). 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.

We attempt to summarize respondents' directly elicited compensatory statements in the fourth column of Table 2 by displaying the percentage of respondents who mentioned each of the feature levels in a compensatory rule. (Respondents could mention one feature level, multiple feature levels, or none.) For example, more than half (60.1%) the respondents mentioned Nokia. Large percentages of respondents also mentioned high camera resolutions. The percentage of compensatory mentions from direct elicitation is significantly correlated with the HB logit, additive utility posterior mean partworths ($\rho = .72$, $p < .001$). The posterior means of the partworths are also significantly negatively correlated with the directly elicited feature-elimination percentages ($\rho = –.49$, $p < .02$). In our application, the directly elicited compensatory percentages are significantly negatively correlated with directly elicited elimination percentages ($\rho = –.71$, $p < .001$).

We can also summarize the output of UDE by addressing a specific managerial question. For example, if Lenovo was considering launching a HK$2500, pink, small-screen, thick, rotational phone with a .5 MP camera resolution, the majority of respondents (67.8%) would not even consider it. In contrast, almost everyone (all but 7.7%) would consider a Nokia, HK$2000, silver, large-screen, slim, slide phone with 3.0 MP camera resolution. Alternatively, we could use respondent-level direct-elicitation data to identify market segments (an analogy to what is now done with respondent-level partworth posterior means).

### *SCALABILITY: THE AUTOMOTIVE STUDY*

To test scalability, we select a product category and set of features that strains (or makes infeasible) decomposition. For the noncompensatory decomposition approaches in Table 1, running time increases exponentially with the number of feature levels (53 feature levels in automobiles versus 24 in mobile phones), requiring a computational factor on the order of 500 million. An HB additive logit model is feasible but strained. Limits on respondent attention suggest that we can measure consideration for, at best, far fewer profiles than would be required by a D-efficient orthogonal design (25,600 profiles in the automotive study). Automotive industry experience suggests that approximately 30 profiles can be evaluated in a comparative study.

The mobile phone study implies that a UDE task might be better than an SDE task. However, this may be the result of the particular structured task tested. Thus, we include an alternative, widely applied, structured task, CASEMAP, which collects self-explicated data on both elimination and compensatory decision rules. CASEMAP has the additional advantage of not requiring the qualitative data to be coded. We expect both the UDE e-mail task and the SDE CASEMAP to scale to a realistic automotive experimental design.

We draw on an experimental design a major U.S. automaker uses to develop strategies to increase consideration of its vehicles (Dzyabura and Hauser 2010). We used pretests to modify the feature levels for a student sample (versus a national panel of auto intenders). In total, 204 students at a U.S. university completed the study. The $20 \times 7 \times 5^2 \times 4 \times 3^4 \times 2^2$ design was as follows:

- •Brand: Audi, BMW, Buick, Cadillac, Chevrolet, Chrysler, Dodge, Ford, Honda, Hyundai, Jeep, Kia, Lexus, Mazda, Mini, Nissan, Scion, Subaru, Toyota, and Volkswagen
- •Body type: compact sedan, compact sport-utility vehicle (SUV), crossover, hatchback, midsize SUV, sports car, and standard sedan
- •EPA mileage: 15, 20, 25, 30, and 35 miles per gallon
- •Glass package: none, defogger, sunroof, and both
- •Transmission: standard, automatic, and shiftable automatic
- •Trim level: base, upgrade, and premium
- •Quality of workmanship rating: Q3, Q4, and Q5
- •Crash test rating: C3, C4, and C5
- •Power seat: yes and no
- •Engine: hybrid and internal combustion
- •Price: profile prices, which varied from $16,000 to $40,000, based on five manipulated levels plus feature-based prices

Opening screens explained all features to respondents using both text and pictures. As training in the features, respondents evaluated a small number of warm-up profiles. We used icons and short text descriptions of the features throughout the online survey, and respondents could return to explanation screens at any time with a single click. Pretests indicated that respondents understood the feature descriptions well.

### *Respondent Tasks for the Automotive Study*

We modified the e-mail UDE task and the three-panel decomposition task to address automobiles rather than mobile phones. For decomposition, we chose 30 automobile profiles randomly from the orthogonal design, eliminating unrealistic profiles such as a Mini Cooper SUV. (We redrew profiles for every respondent with a resulting D-efficiency of .98.) We programmed the CASEMAP task to mimic as closely as possible the descriptions in the work of Srinivasan (1988) and Srinivasan and Wyner (1988). Respondents indicated unacceptable feature levels, indicated their most- and least-preferred level for each feature, identified the most important critical feature, rated the importance of every other feature relative to the critical feature, and scaled preferences for levels within each feature. (We defined importance as the relative value of moving from the least-preferred level to the most-preferred level.)

After general instructions, an introduction to the features and levels, a description of the incentives, and warm-up questions, respondents completed each of the three tasks. We randomized the order of the tasks to mitigate the impact of order effects, if any, on relative comparisons among methods. (For screen shots of the task, see the Web Appendix at http://www.marketingpower.com/jmrfeb11.) Because of the length of the automotive survey and from the results of the mobile phone study, we did not include an initial validation task but rather relied on the delayed task. The delayed validation task used the same format as the decompositional task, drawing 30 profiles per respondent randomly from a second orthogonal design (D-efficiency = .98).

We pretested all instructions, tasks, feature levels, and incentives with 34 respondents. At the end of the pretests, respondents indicated that they understood all tasks, feature levels, and incentives. Respondents were blind to the hypotheses of the study.

### *Incentives*

We structured the incentives in the automotive study the same way as in the mobile phone study, with one key exception: It was not feasible to guarantee $40,000 for an automobile plus cash to 1 of every 30 respondents. To address this problem, we bought prize indemnity insurance. For a fixed fee, we were able to offer to a chosen respondent a reasonable chance that he or she would get $40,000 toward an automobile (plus cash). The features and price (≤$40,000) would be determined by the respondent's answers to one of the four sections of the survey (three calibration tasks and one validation task). Specifically, respondents were told that one randomly selected respondent would draw 2 of 20 envelopes. If both envelopes contained a winning card, the respondent won the $40,000 prize.[5] This is a standard procedure in drawings of this type. Such drawings are common for radio or automotive promotions. Pretests indicated that these incentives were sufficient to motivate respondents to think hard and provide truthful answers. In addition, all respondents received a fixed incentive of $15 when they completed both the initial and the delayed questionnaires.

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." Although the task and the incentives were easiest to understand for CASEMAP ($p < .05$), they seemed to be easy to understand for all three methods. We also asked the participants whether the tasks "enable them to accurately express their preferences." Respondents believed that the UDE and CASEMAP tasks enabled them to express their preferences more accurately than the decompositional task ($p < .01$), with no significant difference between UDE and CASEMAP tasks. In general, respondents enjoyed the three tasks, found them easy to do, put more effort into the tasks because of the incentives, and found the pictures helpful; however, they believed that the tasks took a fair amount of time. (For more details, see the Web Appendix at http://www.marketingpower.com/jmrfeb11.)

### *Results of the Automotive Study*

Table 3 reports the rescaled KL divergence for the three rotated methods and for the null models. Because RMSE

[5] In the actual drawing, the first, but not the second, envelope was a winning envelope. Because the $40,000 prize required that both envelopes be winning envelopes, the respondent received the $200 consolation prize.

relies on consideration shares among profiles in the validation data, we could not calculate it for the automotive data, in which the 25,600 orthogonal profiles are spread sparsely among the 204 respondents.

Table 3 suggests that all three methods predict better than either null model. We find that UDE predicts consideration sets better than decompositional HB logit models, reflecting the difficulty in obtaining data for decompositional methods in complex product categories. Of the two direct-elicitation methods, the UDE task (e-mail) seems to predict consideration better than the SDE task (CASEMAP). This is consistent with the mobile phone study (unstructured > structured). It is also consistent with the hypothesis that respondents' heuristic rules for consideration are cognitively simple and that SDE encourages respondents to overstate elimination rules. For example, CASEMAP-based rules miss considered profiles significantly more than UDE ($p < .001$) or decomposition ($p < .001$).

### *Training Effects*

In the automotive data, we randomized task order. "Training" occurs if the task followed at least one other task. (There were no significant effects between second and third.) The results show that UDE benefits from training but remains best regardless of whether training occurred. In particular, we note the following:

- •With training, UDE is significantly better than both CASEMAP and decomposition. ($p < .001$, KL = 16.1% versus 8.2% and 6.5%, respectively).
- •Without training, UDE is better than both CASEMAP and decomposition but not significantly so ($p > .05$, 8.4% versus 6.8% and 6.9%, respectively).
- •Training benefits UDE significantly ($p < .002$, KL = 16.1% versus 8.4%).
- •Training does not benefit either decomposition or CASEMAP significantly ($p > .05$, KL = 8.2% versus 6.8% and KL = 6.5% versus 6.8%, respectively).

Table 3

PREDICTIVE ABILITY AUTOMOBILE STUDY

| | *Delayed Validation Relative KL Divergence*[a] *(%)* |
|---|---|
| *Decompositional Methods*[b] | |
| HB logit, additive utility | 6.6 |
| HB logit, q-compensatory | 3.7 |
| *CASEMAP (Version of SDE)* | |
| Match cutoff | 7.8 |
| Estimated cutoff[c] | 7.4 |
| *UDE* | |
| Match cutoff | 13.6* |
| Estimated cutoff[d] | 13.2* |
| *Null Models* | |
| Reject all | .0 |
| Random proportional to consideration share in calibration data | .0 |

*Best in column or not significantly different from best at the .05 level.

[a] Rescaled, such that larger is better.

[b] Greedoid dynamic program and logical analysis of data are not computationally feasible for the automotive study.

[c] We determined the utility cutoff in calibration data and then applied it to validation data.

[d] Logit-based estimation of consideration set size as in the mobile phone study.

Training seems to effect a significant improvement in UDE, almost doubling the KL percentage. We observe the training effect for challenging initial tasks that cause respondents to think deeply about their decision process (CASEMAP or a 30-profile evaluation). This training is substantial even though the questionnaire began with a few-profile warm-up exercise. Perhaps further research will be able to untangle why the training effect is much stronger for UDE than for the other methods. (It is possible that a larger sample identifies a significant training effect for the other methods.) In summary, the automotive data suggest that UDE scales to complex product categories better than SDE or decompositional methods, it is feasible to provide realistic incentives even for expensive durable goods, and there is a substantial training effect for UDE.

## *PROMISE AND CHALLENGES*

### *Promise*

Together, the mobile phone and automotive studies suggest that UDE holds promise for further development. Discussions with market research managers with expertise in both quantitative and qualitative methods suggest that for typical sample sizes, the cost of UDE is comparable to that for decompositional methods and structured self-explication. Although UDE requires independent coders, such coders are often paid at lower rates than experienced quantitative analysts. Many market research firms have experienced, trained coders to handle qualitative data, but these coders lack the same depth of experience for advanced statistics (though the widely available Sawtooth software helps). If the results in this article generalize, it seems that the choice of decomposition or UDE for modest experimental designs should be made on grounds other than predictive ability or cost. For complex experimental designs, UDE may be more feasible than decomposition. (For extremely large sample sizes, UDE may become too expensive.)

A concern might be that the e-mail format could prove cumbersome if there were even more features than in the automotive study. Although this has 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, and Johnson 1993; Shugan 1980). It is reasonable to hypothesize that such heuristic decision processes can be captured in an e-mail/narrative format. With UDE, respondents need only describe rules for the feature levels they use to evaluate profiles. If the decision rules are simple, the number of elicited features or feature levels will be small.

One final advantage of UDE is the serendipitous insights that come naturally with qualitative data. In comparison, decompositional methods require additional qualitative questions and the requisite coding. For example, some mobile phone respondents gave reasons for their decision rules such as "rotational phones tend to break down" or "Lenovo has a younger image."

### *Challenges*

The mobile phone and automotive studies are proof of concept, but many challenges remain, such as the following:

1. *Training*: UDE benefits from training more than CASEMAP and decomposition even though the validation occurred a

week after the tasks. Fortunately, the automotive study suggests that respondents can complete both a training task and a 30-profile UDE task with reasonable incentives. Further research could untangle whether respondents are learning the task or learning their own decision rules (Payne, Bettman, and Johnson 1988, 1993). Initial results suggest that UDE applications include a substantial training task before prompting respondents to compose the e-mail. CASEMAP or 30-profile evaluation was sufficient, but there might be other tasks that are more efficient.

2. *Consideration set size*: UDE predictions benefit from a calibrated model of consideration set size. In our applications, we used data from profile evaluations, but other tasks might be more efficient. Until more efficient tasks are tested, the need for a consideration set size model partially mitigates the value of UDE for modest-sized designs (though its value remains for complex designs). Efficient tasks could serve the dual role of calibration and training even if the decomposition data are not otherwise analyzed.
3. *Big-ticket business-to-business products*: We have not yet tested whether incentive alignment can be extended to big-ticket business-to-business products. Researchers could try prize indemnity insurance for business-to-business products if the firm has already solved the agency problem so that its employees act in the best interests of the firm.
4. *Incentives for consideration decisions*: 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. Even the definition of consideration is an open debate (Brown and Wildt 1992). Our incentives seem to have internal validity, motivate respondents to think hard and accurately, and are easy to understand, but they could be improved with further experimentation and experience. We would retain the prize, the dispute resolution among agents, and the agent-auditing process but experiment with different wordings and/or award procedures.
5. *Improved coding procedures*: To provide a conservative test, we wanted to minimize subjectivity in the coding. This is both a disadvantage, because we rely on human judgment, and a potential opportunity if more aggressive coding procedures can be developed to further mine the compensatory statements in the qualitative data.
6. *Alternative benchmarks*: Although we attempted to choose a reasonably complete set of benchmarks for the consider-versus-not-consider task, testing versus other benchmarks could yield further insights. We might also improve direct elicitation with adaptive self-explication (e.g., Netzer and Srinivasan 2011). We could obtain more efficient profile evaluations with methods based on adaptive learning and belief propagation (Dzyabura and Hauser 2010), and HB methods could replace machine-learning noncompensatory estimation.
7. *Managerial summaries*: There are challenges in finding efficient ways to summarize the managerial outputs of noncompensatory decision rules, whether they are from direct elicitation or decomposition.

Many other open questions remain, such as degrees of 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 (for which respondents might be more likely to use noncompensatory heuristics).

## 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), 187–97.

Bateson, John E.G., David Reibstein, and William Boulding (1987), "Conjoint Analysis Reliability and Validity: A Framework for Future Research," in *Review of Marketing*, Michael Houston, ed. Chicago: American Marketing Association, 451–81.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak (1964), "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, 9 (July), 226–32.

Boros, Endre, Peter L. Hammer, Toshihide Ibaraki, and Alexander Kogan (1997), "Logical Analysis of Numerical Data," *Mathematical Programming*, 79 (1–3), 163–90.

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), 1332–46.

Brown, Juanita J. and Albert R. Wildt (1992), "Consideration Set Measurement," *Journal of the Academy of Marketing Science*, 20 (3), 235–63.

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

Ding, Min (2007), "An Incentive-Aligned Mechanism for Conjoint Analysis," *Journal of Marketing Research*, 44 (May), 214–23.

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

———, Young-Hoon Park, and Eric T. Bradlow (2009), "Barter Markets for Conjoint Analysis," *Management Science*, 55 (6), 1003–1017.

Dzyabura, Daria and John R. Hauser (2010), "Active Learning for Consideration Heuristics," working paper, MIT Sloan School of Management, Massachusetts Institute of Technology.

Elrod, Terry (2001), "Recommendations for Validation of Choice Models," in *2001 Sawtooth Conference Proceedings*. Sequim, WA: Sawtooth Software, 225–43.

———, Jordan Louviere, and Krishnakumar S. Davey (1992), "An Empirical Comparison of Ratings-Based and Choice-Based Conjoint Models," *Journal of Marketing Research*, 29 (August), 368–77.

Fishbein, Martin 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), [available at 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), 494–509.

Gigerenzer, Gerd and Daniel G. Goldstein (1996), "Reasoning the Fast and Frugal Way: Models of Bounded Rationality," *Psychological Review*, 103 (4), 650–69.

Green, Paul E. (1984), "Hybrid Models for Conjoint Analysis: An Expository Review," *Journal of Marketing Research*, 21 (May), 155–69.

——— and Kristiaan Helsen (1989), "Cross-Validation Assessment of Alternatives to Individual-Level Conjoint Analysis: A Case Study," *Journal of Marketing Research*, 26 (August), 346–50.

———, ———, and Bruce Shandler (1988), "Conjoint Internal Validity Under Alternative Profile Presentations," *Journal of Consumer Research*, 15 (December), 392–97.

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

Griffin, Abbie and John R. Hauser (1993), "The Voice of the Customer," *Marketing Science*, 12 (1), 1–27.

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

———, Olivier Toubia, Theodoros Evgeniou, Daria Dzyabura, and Rene Befurt (2011), "Cognitive Simplicity and Consideration Sets," *Journal of Marketing Research*, 48, forthcoming.

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

——— and Kenneth J. Wisniewski (1982), "Dynamic Analysis of Consumer Response to Marketing Strategies," *Management Science*, 28 (5), 455–86.

Hoepfl, Robert T. and George P. Huber (1970), "A Study of Self-Explicated Utility Models," *Behavioral Science*, 15 (5), 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), 1860–72.

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*, 30 (February), 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), 185–95.

Jedidi, Kamel and Rajeev Kohli (2005), "Probabilistic Subset-Conjunctive Models for Heterogeneous Consumers," *Journal of Marketing Research*, 42 (November), 483–94.

Johnson, Eric J., Robert J. Meyer, and Sanjoy Ghose (1989), "When Choice Models Fail: Compensatory Models in Negatively Correlated Environments," *Journal of Marketing Research*, 26 (August), 255–70.

Klein, Noreen M. (1986), "Assessing Unacceptable Attribute Levels in Conjoint Analysis," in *Advances in Consumer Research*, Vol. 14, M. Wallendorf and P. Anderson, eds. Provo, UT: Association for Consumer Research, 154–58.

Kohli, Rajeev and Kamel Jedidi (2007), "Representation and Inference of Lexicographic Preference Models and Their Variants," *Marketing Science*, 26 (3), 380–99.

Kramer, Thomas (2007), "The Effect of Measurement Task Transparency on Preference Construction and Evaluations of Personalized Recommendations," *Journal of Marketing Research*, 44 (May), 224–33.

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

Kullback, Solomon and Richard A. Leibler (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22 (1), 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*, 21 (November), 456–62.

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

Martignon, Laura and Ulrich Hoffrage (2002), "Fast, Frugal, and Fit: Simple Heuristics for Paired Comparisons," *Theory and Decision*, 52 (1), 29–71.

Moore, William L. and Ekaterina Karniouchina (2006), "Screening Rules and Consumer Choice: A Comparison of Compensatory vs. Non-Compensatory Models," working paper, David Eccles School of Business, University of 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*, 16 (3), 261–74.

Netzer, Oded and V. Srinivasan (2011), "Adaptive Self-Explication of Multiattribute Preferences," *Journal of Marketing Research*, 48 (February), 140–56.

Olshavsky, Richard W. and Franklin Acito (1980), "An Information Processing Probe into Conjoint Analysis," *Decision Sciences*, 11 (July), 451–70.

Park, Young-Hoon, Min Ding, and Vithala R. Rao (2008), "Eliciting Preference for Complex Products: Web-Based Upgrading Method," *Journal of Marketing Research*, 45 (October), 562–74.

Payne, John W. (1976), "Task Complexity and Contingent Processing in Decision Making: An Information Search," *Organizational Behavior and Human Performance*, 16 (2), 366–87.

———, James R. Bettman, and Eric J. Johnson (1988), "Adaptive Strategy Selection in Decision Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14 (3), 534–52.

———, ———, 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), 135–48.

Prelec, Dražen (2004), "A Bayesian Truth Serum for Subjective Data," *Science*, 306 (October 15), 462–66.

Roberts, John H. and James M. Lattin (1991), "Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 28 (November), 429–40.

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

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

——— (2004), "The CBC Hierarchical Bayes Technical Paper," research report, Sawtooth Software.

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

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*, 19 (2), 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), 286–91.

——— and Gordon A. Wyner (1988), "CASEMAP: Computer-Assisted Self-Explication of Multiattributed Preferences," in *Handbook on New Product Development and Testing*, W. Henry, M. Menasco, and K. Takada, eds. 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), 679–97.

Toubia, Olivier (2006), "Idea Generation, Creativity, and Incentives," *Marketing Science*, 25 (5), 411–25.

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

———, ———, and Duncan Simester (2004), "Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis," *Journal of Marketing Research*, 41 (February), 116–31.

———, 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), 428–41.

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), 532–49.