

New Approaches to Idea Generation and Consumer Input in the Product Development Process

By
Olivier Toubia


Ingénieur, Ecole Centrale Paris, 2000
M.S. Operations Research, Massachusetts Institute of Technology, 2001


SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF


DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

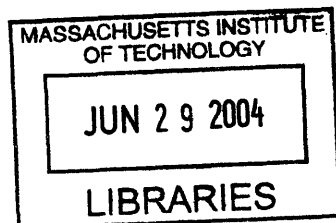
JUNE 2004

© Massachusetts Institute of Technology. All rights reserved.

Signature of Author: _____ 
Sloan School of Management
May 10th, 2004

Certified by: _____  *May 10, 2004*
John R. Hauser
Kirin Professor of Marketing
Thesis Supervisor

Accepted by: _____ 
Birger Wernerfelt
Professor of Management Science
Chair, Sloan Doctoral Program



ARCHIVES :

New Approaches to Idea Generation and Consumer Input in the Product Development Process

By
Olivier Toubia

Submitted to the Sloan School of Management on May 10th, 2004 in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Management Science

ABSTRACT

This thesis consists of five related essays which explore new approaches to help design successful and profitable new products. The primary focus is the front end of the process where the product development team is seeking improved input from customers and improved ideas for developing products based on that input. Essay 1 examines whether carefully tailored ideation incentives can improve creative output. The influence of incentives on idea generation is studied using a formal model of the ideation process. A practical, web-based, asynchronous “ideation game” is developed, allowing the implementation and test of various incentive schemes. Using this system, an experiment is run, which demonstrates that incentives do have the capability to improve idea generation, confirms the prediction from the theoretical analysis, and provides additional insight on the mechanisms of ideation. Essay 2 proposes and tests new adaptive question design and estimation algorithms for partial-profile conjoint analysis. The methods are based on the identification and characterization of the set of parameters consistent with a respondent’s answers. This feasible set is a polyhedron defined by equality constraints, each paired-comparison question yielding a new constraint. Polyhedral question design attempts to reduce the feasible set of parameters as rapidly as possible. Analytic Center estimation relies on the center of the feasible set. The proposed methods are evaluated relative to established benchmarks using simulations, as well as a field test with 330 respondents. Essay 3 introduces polyhedral methods for choice-based conjoint analysis, and generalizes the concept of D-efficiency to individual adaptation. The performance of the methods is evaluated using simulations, and an empirical application to the design of executive education programs is described. Essay 4 generalizes the existing polyhedral methods for adaptive choice-based conjoint analysis by taking response error into account in the adaptive design and estimation of choice-based polyhedral questionnaires. The validity of the proposed approach is tested using simulations. Essay 5 studies the impact of Utility Balance on efficiency and bias. A new definition of efficiency (M-efficiency) is also introduced, which recognizes the necessity to match preference questions with the quantities used in the ultimate managerial decisions.

Thesis Supervisor: John R. Hauser
Title: Kirin Professor of Marketing

Acknowledgments

When I first started being a student at MIT in September 1999, my goal was to quickly get a Master's in Operations Research. I had just gotten engaged. Today, on May 9th 2004, I am writing the acknowledgment section of my doctoral dissertation in marketing. My son Eitan has just wished my wife Alexandra her first mother's day. What happened? One hypothesis is that I have been brainwashed and manipulated. Another hypothesis is that I have had the immense privilege of being surrounded with outstanding people over the past five years. Let us focus on the second hypothesis. Actually, it started well before September 1999. My parents have always provided me with love, support, and advice. They have taught me important values and helped me develop myself, without trying to force me into any predefined path. The well being of their children (and now, grandchildren) has always been their first priority and they have always supported our decisions "as long as you're healthy and happy and if that's what you want". I hope to always remember their lesson that taking care of others can be more rewarding than taking care of oneself. My late maternal grandfather, "Popal", has been my greatest role model. I have always admired him, his strength, his wisdom, his willpower and his perspicacity. I hope and I believe that he would have approved the career that I have adopted, and I like to think that he might have made similar decisions in similar situations. My late maternal grandmother, "Mamounie", gave me her unconditional love and affection, and transmitted to me the value of academic achievement. Their memories have been with me all along, and will continue to guide my professional and personal lives. My paternal grandparents "Mammy" and "Daddy" have been a model of hope and optimism, have taught me how to love and cherish life, and have shown me the importance of family. My brother Didier and my sister Valérie have excelled at the difficult job of having me as a little brother, giving me advice and leading the way for me. Then, of course, came Alexandra, who supported my decision to go study abroad and with whom I have shared all the exciting, happy and uncertain moments of this adventure. She has been my wife, my friend, my companion and my supporter. I would have never been able to have a balanced life at MIT without her. For me she gave up her hope for a predictable life, and we are now ready to share whatever will come next. Finally, my cousin Isabelle and her family Philippe, Nicolas and Lionel have been our family away from home. It has been great to share holidays with them, and to see Nicolas and Lionel grow.

At MIT, I have had the privilege to meet and work with some incredibly brilliant professors and fellow students. First, of course, my advisor, John Hauser, has been a perfect "academic father". John has taught me how to apply mathematical tools to the study of marketing phenomena, how to balance scientific inquiry with managerial relevance, how to apply the History of science to analyze the worth of a research project, and how to expose research. His perfectionism, his extremely high personal standards, and his dedication to hard work have been inspiring. He has been a mentor for me at all levels, training me on research, teaching, and on how to be a good husband and father. He has also been incredibly human, reassuring me in the most troubled times. Duncan Simester has continuously amazed me with his ability to understand, analyze and critique virtually any intellectual argument. He has taught me how important it is to appropriately position a research project and to isolate the essence of an idea. Ely Dahan has been one of the most influential persons in my decision to pursue a PhD. Beyond a colleague, he has been a friend and we will never forget this 4th of July weekend with him and his family in Hyannis. I would also like to thank Dan Ariely for his benevolent feedback on my research, and for his generous funding. He has been an example of dedication to research, and his PhD seminars fed not

only my body, but also my brain. Drazen Prelec has supported me and advised me on my job market paper. I admire his calm and the deepness of his thoughts. I have also been fortunate to interact with John Little, who confirmed my hypothesis that the most brilliant and accomplished researchers are often also the most friendly and caring human beings. Glen Urban also confirmed this hypothesis. He introduced me to teaching and gave me an opportunity to better prepare for my next job. Birger Wernerfelt, by exposing me to his research paradigm, has forced me to challenge my own and, I believe, to improve the rigor and the scientific value of my research. JinGyo Kim introduced me to the world of hierarchical Bayes and helped me acquire some indispensable tools. I hope that Shane Frederick made me a slightly smarter person, by asking all those I.Q. questions at the lunch table.

Rosa Blackwood gave me some advice and support, often working extra time for me although she does not even work for me! The marketing group would not be the same without her good humor and her great cakes. Sandra Crawford-Jenkins shared with me her music and her kindness.

I will also surely miss my fellow students. My office mate Wei Wu could not have been more pleasant. My former office mate Robert Zeithammer made a little room for me in his office and guided me through my first year. On Amir has been like a big brother, giving me some advice and sharing his wisdom with me. Jiwoong Shin has gone through the job market with me and offered me some guidance while studying for the general exams. Leonard Lee has been a lot of fun to be around and to study with. Nina Mazar has brought her sophisticated and delicate touch to the group. Kristina Shampan'er has contributed her wit and critical thinking. Ray Weaver, with his great sense of humor, his MBA and his two children, has not been a typical first year student. I hope that he will graduate before Shane turns him into an addicted gambler.

I would like to conclude by thanking a very special person, who in just three months has changed my life forever and introduced me to the joys of fatherhood. I am proud to give to my son Eitan Toubia his first Google hit.

Table of Contents

Acknowledgments.....	3
Chapter 1: Idea Generation, Creativity, and Incentives.....	7
1. Introduction.....	8
2. Theoretical Analysis	9
3. An Ideation Game	20
4. Experiment.....	24
5. Summary and Future Research	32
References.....	34
Appendix : proofs of the propositions	37
Chapter 2: Fast Polyhedral Adaptive Conjoint Estimation.....	46
1. Polyhedral Methods for Conjoint Analysis.....	47
2. Polyhedral Question Design and Estimation	49
3. Monte Carlo Simulations	60
4. Results of the Initial Monte Carlo Experiments	64
5. The Role of Self-Explicated Questions.....	69
6. Empirical Application and Test of Polyhedral Methods.....	72
7. Results of the Field Test	80
8. Product Launch	86
9. Conclusions and Future Research.....	86
References.....	91
Appendix 1: Mathematics of Fast Polyhedral Adaptive Conjoint Estimation.....	97
Appendix 2: Internal Validity Tests for Laptop Computer Bags.....	102
Chapter 3: Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis....	103
1. Introduction.....	104
2. Existing CBC Question Design and Estimation Methods	106
3. Polyhedral Question Design Methods	108
4. Monte Carlo Experiments.....	119
5. Application to the Design of Executive Education Programs	127
6. Conclusions and Research Opportunities	130

Endnotes.....	132
References.....	134
Appendix: Mathematics of Polyhedral Methods for CBC Analysis.....	137
Chapter 4: Non-Deterministic Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis.....	140
1. Introduction.....	141
2. Basic polyhedral methods for Choice-Based Conjoint Analysis.....	141
3. Taking response error into account.....	145
4. Simulations	152
5. Conclusions and Future Research.....	157
References.....	158
Chapter 5: Properties of Preference Questions: Utility Balance, Choice Balance, Configurators, and M-Efficiency.....	159
1. Motivation.....	160
2. Efficiency in Question Design.....	163
3. Utility Balance for the Metric Paired-Comparison Format	164
4. Choice Balance for the Stated-Choice Format.....	169
5. Configurators	174
6. Generalizations of Question Focus: M-Efficiency	177
7. Summary and Future Research	182
References.....	184
Appendix.....	187

Chapter 1: Idea Generation, Creativity, and Incentives

Abstract

Idea generation (ideation) is critical to the design and marketing of new products, to marketing strategy, and to the creation of effective advertising copy. However, there has been relatively little formal research on the underlying incentives with which to encourage participants to focus their energies on relevant and novel ideas. Several problems have been identified with traditional ideation methods. For example, participants often free ride on other participants' efforts because rewards are typically based on the group-level output of ideation sessions.

This paper examines whether carefully tailored ideation incentives can improve creative output. I begin by studying the influence of incentives on idea generation using a formal model of the ideation process. This model identifies conditions under which it is efficient to simply reward participants based on their contributions, and other conditions under which it is not. I show that in the latter case, the group's output can be improved by rewarding participants for their impact on the other participants' contribution. I then develop a practical, web-based, asynchronous "ideation game," which allows the implementation and test of various incentive schemes. Using this system, I run an experiment, which demonstrates that incentives do have the capability to improve idea generation, confirms the predictions from the theoretical analysis, and provides additional insight on the mechanisms of ideation.

1. Introduction

Idea generation (ideation) is critical to the design and marketing of new products, to marketing strategy, and to the creation of effective advertising copy. In new product development, for example, idea generation is a key component of the front end of the process, often called the “fuzzy front end” and recognized as one of the highest leverage points for a firm (Dahan and Hauser 2001).

The best known idea generation methods have evolved from “brainstorming,” developed by Osborn in the 1950’s (Osborn 1957). However, dozens of studies have demonstrated that groups generating ideas using traditional brainstorming are less effective than individuals working alone (see Diehl and Stroebe 1987 or Lamm and Trommsdorff 1973 for a review). One cause identified for this poor performance is free riding (Williams et al. 1981, Kerr and Bruun 1983, Harkins and Petty 1982). In particular, participants free ride on each other’s creative effort because the output of idea generation sessions is typically considered at the group level, and participants are not rewarded for their individual contributions. This suggests that idea generation could be improved by providing appropriate incentives to the participants. Note that this free riding effect is likely to be magnified as firms seek input from customers (who typically do not have a strong interest in the firm’s success) at an increasingly early stage of the new product development process (von Hippel and Katz 2002).

Surprisingly, agency theory appears to have paid little attention to the influence of incentives on agents’ *creative* output. On the other hand, classic research in social psychology suggests that incentives might actually have a negative effect on ideation. For example, the Hull-Spence theory (Spence 1956) predicts an enhancing effect of rewards on performance in simple tasks but a detrimental effect in complex tasks. The reason is that rewards increase indiscriminately all response tendencies. If complex tasks are defined as tasks in which there is a predisposition to make more errors than correct responses, then this predisposition is amplified by rewards. Similarly, the social facilitation paradigm (Zajonc 1965) suggests that, insofar as incentives are a source of arousal, they should enhance the emission of dominant, well-learned responses, but inhibit new responses, leading people to perform well only at tasks with which they are familiar. McCullers (1978) points out that incentives enhance performance when it relies on making “simple, routine, unchanging responses,” (p. 14) but that the role of incentives is far less

clear in situations that depend heavily on flexibility, conceptual and perceptual openness, or creativity. McGraw (1978) identifies two conditions under which incentives will have a detrimental effect on performance: “first, when the task is interesting enough for subjects that the offer of incentives is a superfluous source of motivation; second, when the solution to the task is open-ended enough that the steps leading to a solution are not immediately obvious.” (p. 34)

In this paper I examine whether carefully tailored ideation incentives can improve creative output. In Section 2, I study the influence of incentives on idea generation using a formal model of the ideation process. In Section 3, as a step towards implementing and testing the conclusions from Section 2, I present a practical, web-based, asynchronous “ideation game,” which allows the implementation and test of various incentive schemes. Finally, in Section 4, I describe an initial experiment conducted using this system. This experiment demonstrates that incentives do have the capability to improve idea generation, is consistent with the predictions from the theoretical analysis, and provides additional insight on the mechanisms of ideation. Section 5 concludes and suggests some possibilities for future research.

2. Theoretical Analysis

The goal of this section is to develop a theoretical framework allowing the study of idea generation incentives. One characteristic of ideas is that they are usually not independent, but rather build on each other to form streams of ideas. This is analogous to academic research, in which papers build on each other, cite each other, and constitute streams of research. Let us define the contribution of a stream of ideas as the moderator’s valuation for these ideas (for simplicity I shall not distinguish between the moderator and his or her client who uses and values the ideas). The expected contribution of a stream of ideas can be viewed as a function of the number of ideas in that stream. Let us assume that the expected contribution of a stream of n ideas is a non-decreasing, concave function of n , such that each new idea has a non-negative expected marginal contribution and such that there are diminishing returns to new ideas within a stream. Note that the analysis would generalize to the case of increasing, or inverted U-shaped, returns. (This might characterize situations in which the first few ideas in the stream build some foundations, allowing subsequent ideas to carry the greatest part of the contribution.)

The model proposed here is based on Kuhn’s notion of an “essential tension” between convergent thinking and divergent thinking (Kuhn 1977, Simonton 1988). Divergent thinking

involves changing perspectives and trying new approaches to problems, while convergent thinking relies on linear and logical steps and tends to be more incremental. Kuhn claims that “since these two modes of thought are inevitably in conflict, it will follow that the ability to support a tension that can occasionally become almost unbearable is one of the prime requisites for the very best of scientific research.” (p. 226)

Although Kuhn’s focus is the history of science, his insights apply to idea generation. If several streams of ideas have been established in an idea generation session, a participant has a choice between contributing to one of the existing streams, and starting a new stream. If he or she decides to contribute to an existing stream, he or she might have to choose between contributing to a longer, more mature stream, and a newer, emerging one. If we assume that each stream of ideas reflects a different approach to the topic of the session, then the problem faced by the participant can be viewed equivalently as that of deciding which approach to follow in his or her search for new ideas.

I assume that some approaches can be more fruitful than others, and that the participant does not know a priori the value of each approach but rather learns it through experience. He or she then has to fulfill two potentially contradictory goals, reflecting Kuhn’s essential tension: increasing short-term contribution by following an approach that has already proven fruitful, versus improving long-term contribution by investigating an approach on which little experience has been accumulated so far.

For example, let us consider a group of participants generating ideas on “How to improve the impact of the UN Security Council?” (which was the topic of the experimental sessions reported in Section 4). Let us assume that at one point in the session, two streams of ideas have emerged: a long stream, composed of ten ideas, all approaching the topic from a legal standpoint, and a shorter stream, with only two ideas, both approaching the topic from a monetary standpoint. Each participant then has to decide whether to search for new ideas following a legal, or a monetary approach to the topic. The legal approach has already proven fruitful, whereas the monetary approach is more risky but potentially more promising.

This problem is typical in decision making and is well captured by a representation known as the multi-armed bandit model (Bellman 1961)¹. For simplicity, let us first restrict ourselves to a two-armed bandit, representing a choice between two approaches A and B . Let us assume that by spending one unit of time thinking about the topic using approach A (respectively B), a participant generates a relevant idea on the topic with unknown probability p_A (respectively p_B). Participants hold some beliefs on these probabilities, which are updated after each trial. Note that starting a new stream of ideas can be viewed as a special case in which a participant tries an approach that has not been used successfully yet.

Let us complement this classical multi-armed bandit model by assuming that when N_A trials of approach A have been made, a proportion \hat{p}_A of which were successful, then the next idea found using approach A has contribution $\alpha^{\hat{p}_A N_A}$ (similarly for approach B). This assumption captures the fact that new ideas in a stream have diminishing marginal contribution ($0 \leq \alpha \leq 1$ by assumption). In the context of our previous example, this assumption implies that the third idea following the monetary approach is likely to have a greater marginal contribution than the eleventh idea following the legal approach, which increases the attractiveness of the riskier monetary approach.

Some problems (like mathematical problems or puzzles) have objective solutions. Once a solution has been proposed, there is usually little to build upon. Such topics would be represented by a low value of α . On the other hand, many problems considered in idea generation are more “open-ended” (trying to improve a product or service) and are such that consecutive ideas refine and improve each other. Such situations correspond to higher values of α . The parameter α could also depend on the preferences of the user(s) of the ideas. For example, the organizer of the session might value many unrelated, rough ideas (low α), or, alternatively, value ideas that lead to a deeper understanding and exploration of certain concepts (high α).

To analyze the tension in ideation, I first identify and characterize three stages in the idea generation process. Next I consider the simple incentive scheme consisting of rewarding participants based on their individual contributions, and show that although it leads to optimal search when α is small enough, it does not when α is large. For large α , this suggests the need for incen-

¹ Bandit is a reference to gambling machines (slot machines) that are often called one-armed bandits. In a multi-armed bandit, the machine has multiple lever (arms) and the gambler can choose to pull any one of them at a time.

tives that interest participants in the future of the group's output. I next show that, although such incentives allow aligning the objectives of the participants with that of the moderator of the session, they might lead to free riding. Avoiding free riding requires further tuning of the rewards and can be addressed by rewarding participants more precisely for the impact of their contributions.

Three stages in the idea generation process

Let us consider a two-period model, in which a single participant searches for ideas, and study the conditions under which it is optimal to follow the approach that has the higher versus the lower prior expected probability of success. For simplicity, I shall refer to the former as the "better-looking approach" and to the latter as the "worse-looking approach."

I assume that the prior beliefs on p_A and p_B at the beginning of period 1 are independent and such that $p_A \sim \text{Beta}(n_{AS}, n_{AF})$ and $p_B \sim \text{Beta}(n_{BS}, n_{BF})$, with $n_{AS}, n_{AF}, n_{BS}, n_{BF} \geq 1$.² The corresponding expected probabilities of success are $\hat{p}_A = E(p_A) = n_{AS}/N_A$; $\hat{p}_B = E(p_B) = n_{BS}/N_B$, where $N_A = n_{AS} + n_{AF}$ and $N_B = n_{BS} + n_{BF}$ can be interpreted as the total number of trials observed for A and B respectively prior to period 1.³ For these assumptions the variance of the beliefs is inversely proportional to the number of trials.

Two effects influence this model. First, if the uncertainty on the worse-looking approach is large enough compared to the uncertainty on the better-looking approach, it might be optimal to choose the worse-looking approach in period 1 although it implies a short-term loss. In particular, if the expected probability of success of the uncertain, worse-looking approach is close enough to that of the more certain, better-looking approach, a success with the worse-looking approach in period 1 might lead to updated beliefs that will make this approach much more appealing than the currently better-looking one.

For example, let us assume that the better-looking approach at the beginning of period 1, B , has a very stable expected probability of success which is around 0.61. By stable we mean that

² These beliefs can be interpreted as corresponding to a situation in which the participant starts with some initial prior uniform beliefs on p_A and p_B ($p_A \sim \text{Beta}(1, 1)$ and $p_B \sim \text{Beta}(1, 1)$) and observes, prior to period 1, $n_{AS}-1$ successes for A , $n_{AF}-1$ failures for A , $n_{BS}-1$ successes for B , and $n_{BF}-1$ failures for B . Because beta priors are conjugates for binomial likelihoods (Gelman et al. 1995), the posterior beliefs after these observations (which are used as priors at the beginning of period 1) are $p_A \sim \text{Beta}(n_{AS}, n_{AF})$ and $p_B \sim \text{Beta}(n_{BS}, n_{BF})$.

³ The uniform prior can be interpreted as resulting from the observation of one success and one failure for each approach.

it will be similar at the beginning of period 2 whether a success or failure is observed for this approach in period 1 (N_B is very large). Let us assume that the worse-looking approach, A , has an expected probability of success of 0.60 at the beginning of period 1. However a success with this approach would increase the posterior probability to 0.67 while a failure would decrease this posterior probability to 0.50.⁴ Then, playing B in period 1 would lead to a total expected contribution of $0.61+0.61=1.22$ (ignoring discounting for simplicity). Playing A in period 1 would lead to an expected total contribution of $0.60+0.60*0.67+0.40*0.61=1.25$ (play A in period 2 if it was successful in period 1), which is higher.

Second, if the number of trials (and successes) on the better-looking approach gets large, this approach becomes over-exploited (due to the decreasing marginal returns) and it might become more attractive to choose the worse looking approach: although the probability of success is lower, the contribution in case of success is higher.

Let us define (we define similar quantities for B):

- $ST(A)$ the expected contribution from period 1 obtained by choosing approach A in period 1 (ST stands for short-term).
- $LT(A)$ the expected contribution from period 2, calculated at the beginning of period 1, if A is played in period 1.

By definition, it is optimal to play B in period 1 if and only if $ST(B)+LT(B) \geq ST(A)+LT(A)$. If B is such that $\hat{p}_B > \hat{p}_A$, then $ST(B) > ST(A)$ unless B has been over-exploited and $ST(B) \leq ST(A)$. The following propositions characterize three stages in the idea generation process, labeled exploration, exploitation, and diversification. If B has been over-exploited, diversification is optimal and A should be played in period 1. When B has not been over-exploited yet, then if α is small enough, the exploitation of B is always optimal. In this case, exploring A would be too costly: at the same time as the participant would learn more about p_A , he or she would also heavily decrease the attractiveness of this approach. On the other hand, when α is large enough, then if A is uncertain enough compared to B , the exploration of A might be optimal in period 1.

⁴ These probabilities would result from $N_A=5$, $n_{AS}=3$, $n_{AF}=2$, implying that $\hat{p}_A = 3/5$ and that \hat{p}_A after a success is updated to $4/6$.

Exploration is characterized by a short-term loss incurred in order to increase long-term contribution.

Proposition 1a: *For all $\hat{p}_B, N_A, \alpha < 1$, for all \hat{p}_A close enough to \hat{p}_B such that $\hat{p}_A < \hat{p}_B$, if B has been over-exploited ($ST(B) \leq ST(A)$), then diversification is optimal in period 1.*

Proposition 1b: *For all \hat{p}_B, N_A , for all α close enough to 0, for all \hat{p}_A close enough to \hat{p}_B such that $\hat{p}_A < \hat{p}_B$, if B has not been over-exploited ($ST(B) > ST(A)$), then the exploitation of B is optimal in period 1.*

Proposition 1c: *For all \hat{p}_B, N_A , for all α close enough to 1 such that $\alpha < 1$, for all \hat{p}_A close enough to \hat{p}_B such that $\hat{p}_A < \hat{p}_B$, there exists $0 < N_B^* < N_B^{**}$ ($0 < V_B^{**} < V_B^*$) such that*

- *The exploitation of B is optimal if $N_B \leq N_B^*$ ($Var(p_B) \geq V_B^*$).*
- *The exploration of A is optimal if $N_B^* \leq N_B \leq N_B^{**}$ ($V_B^{**} \leq Var(p_B) \leq V_B^*$).*
- *Diversification is optimal if $N_B^{**} \leq N_B$ ($Var(p_B) \leq V_B^{**}$).*

The proofs to the propositions are in Appendix 1. Note that the above propositions focus on the more interesting case where the expected probabilities of success of the two approaches are close.

The identification and characterization of these three stages provides a useful background for the study of idea generation incentives. In particular, short-term focus is optimal in the diversification and exploitation stages, such that participants trying to maximize their own contributions behave as if they were maximizing the group's output. In the exploration stage, however, participants have to forego some short-term contribution in order to increase the group's future output. In this case problems arise if participants are not given incentives to internalize the influence of their present actions on the future of the group.

Rewarding participants for their individual contributions

Consider the incentive scheme that consists of rewarding participants based on their individual contributions. In practice, this could be operationalized by having an external judge rate each idea sequentially, and rewarding each participant based on the ratings of his or her ideas. This scheme is probably among the first ones that come to mind, and addresses the free-riding issue mentioned in the introduction. There are also many situations in which individual contribution is likely to be the most sophisticated metric available. It is then useful to identify some conditions under which basing incentives on this metric leads participants to behave in a way that maximizes the group's total contribution, and conditions under which it does not.

Let us introduce a second participant into the model. Assume that, at each period, both players simultaneously choose one of the two approaches. At the end of period 1, the outcomes of both players' searches are observed by both players, and players update their beliefs on p_A and p_B .⁵ Each player's payoff is proportional to his or her individual contribution, i.e., to the sum of the contributions of his or her ideas from the two periods. The contribution of the group is equal to the sum of the contributions of each player. Note that α applies equally for both players, such that if a player finds an idea using a certain approach, the marginal return on this approach is equally decreased for both participants. The assumption that players independently choose which approach to follow and then share the output of their searches is probably more descriptive of electronic ideation sessions (in which participants are seated at different terminals from which they type in their new ideas and read the other participants' ideas) than it is of traditional face-to-face sessions. As will be seen later, electronic sessions have been shown to be more effective than face-to-face sessions (Gallupe et al. 1991, Nunamaker et al. 1987).

Let us consider the subgame-perfect equilibrium (SPE) of the game played by the participants, as well as the socially optimal strategy, i.e., the strategy that maximizes the total expected contribution of the group.

⁵ Equivalently, we could assume that players report the outcome of their $t=1$ search only if it is successful, by submitting a new idea. In this case if a player does not submit an idea at the end of period 1, the other player can infer that her search was unsuccessful. Finally, we could interpret this assumption as meaning that "success" and "failure" correspond to "good" and "fair" ideas, and that both types of ideas are worth submitting.

Low- α case (rapidly decreasing returns)

When α is low, the marginal contribution of successive ideas following the same approach diminishes quickly. In this case, exploration is never optimal, which lowers the value of group interactions. Consequently, actions that are optimal for self-interested participants trying to maximize their own contributions are also optimal for the group. This is captured by the following proposition:

Proposition 2: *For all N_A, N_B , there exists α^* such that $0 < \alpha < \alpha^* \Rightarrow$ for all \hat{p}_A, \hat{p}_B such that $\hat{p}_A < \hat{p}_B$ and \hat{p}_A is close enough to \hat{p}_B , if participants are rewarded based on their own contributions, then a strategy (x, y) is played in period 1 in a SPE of the game iff it is socially optimal.*

High- α case (slowly decreasing returns)

When α is close to 1, there are situations in which exploration is optimal. However, exploration results in a short-term loss suffered only by the explorer, and a potential long-term benefit enjoyed by the whole group (participants are assumed to share their ideas). There are then some cases in which it is socially optimal that at least one player explores in period 1, but no player will do so in any SPE of the game. This is illustrated by the following proposition when $\alpha=1$:

Proposition 3: *If $\alpha=1$, then for all \hat{p}_B , for all $\hat{p}_A < \hat{p}_B$ close enough to \hat{p}_B , there exists $N_{A_{low}} < N_{A_{high}} (V_{low} < V_{high})$ such that if N_B is large enough ($\text{Var}(p_B)$ low enough), then*

- *$N_A < N_{A_{high}} (\text{Var}(p_A) > V_{low})$ implies that it is not socially optimal for both players to choose B in period 1 (at least one player should explore)*
- *$N_{A_{low}} < N_A < N_{A_{high}} (V_{low} < \text{Var}(p_A) < V_{high})$ implies that both players choose B in period 1 in any SPE of the game (no player actually explores)*

Interesting participants in the future of the group's output

Proposition 3 demonstrates a potential misalignment problem when building on ideas matters, that is, when the marginal returns to subsequent ideas are high. In order to overcome this problem, participants should be forced to internalize the effect of their present actions on the future of the other participants' contribution. To study the dynamic nature of the incentives implied by this observation, let us consider a more general framework, with an infinite horizon, N participants, and a per-period discount factor δ . Assume that participant i 's output from period t , y_{it} , has a probability density function $f(y_{it} | x_{1t}, \dots, x_{Nt}, (y_{j\tau})_{j=1 \dots N, \tau=1 \dots t-1})$, where $x_{j\tau}$ is participant j 's action in period τ , and that participant i 's contribution from period t is $C_{it} = C((y_{j\tau})_{j=1 \dots N, \tau=1 \dots t})$. The two-armed bandit model considered earlier is a special case of this general model.

Interesting participants in the group's future output allows aligning the incentives, but it might also lead to free-riding. To illustrate this, let us consider the incentive scheme that consists of rewarding a participant for his or her action in period t according to a weighted average between his or her contribution in period t and the other participants' contribution in period $t+1$.⁶ More precisely, player i receives a payoff for his or her action in period t proportional to $[\gamma \cdot C_{it} + (1 - \gamma) \cdot \delta \cdot \sum_{j \neq i} C_{j,t+1}]$. Let us call $S(\gamma)$ such a scheme. We have the following proposition:

Proposition 4: *In any infinite horizon game with discount factor $\delta < 1$, $S(\gamma)$ aligns the objectives of the moderator and the participants for any form of the contribution and output functions iff $S(\gamma)$ is equivalent to a proportional sharing rule in which player i receives a payoff for his or her actions in period t proportional to $\sum_j C_{jt}$.*

Intuitively, in period t participant i gets, in addition to a share of C_{it} as a reward for his or her actions in period t , a share of the other participants' contribution from period t ($\sum_{j \neq i} C_{j,t}$) as a reward for his or her actions in period $t-1$. However, the fact that this reward is attributed to his or her action in period $t-1$ is irrelevant in period t , since this action has already been taken. Hence

⁶ Proposition 4 would also hold with a weighted average between the participant's contribution in period t and a discounted sum of the other participants' contribution in subsequent periods.

the scheme is equivalent to giving the participant in period t a reward equal to a weighted average between his or her contribution in period t and the other participants' contribution in the same period.

The equivalence to a proportional sharing rule in this example illustrates how group incentives align the objectives at the same time as they introduce free riding (Holmstrom 1982), even in a dynamic setting. In particular, although the objective function of the participants is proportional to that of the moderator, it is possible for a participant to be paid without participating in the search. This imposes a cost on the moderator.

More precisely, let us assume that at each period, players have an option not to search for ideas, or to search at a per-period cost c , and that the first-best level of effort is for all the players to perform a search at each period. Then if the reward given to participant i at period t is equal to $\beta \cdot \sum_j C_{jt}$, β has to be high enough for the incentive compatibility (IC) constraints of the participants to be satisfied (i.e., such that each participant is willing to search in each period). In this case free riding forces the moderator to redistribute a higher share of the created value to the participants.⁷

Rewarding Participants for their Impact

In order to address free riding while still aligning the objectives, one option is to reward each participant more precisely for that part of the group's future contribution that depends directly on his or her actions, i.e., for his or her impact on the group. In particular, the following proposition considers a modification of the above scheme that rewards participant i for his or her action in period t based on a weighted average between his or her contribution in period t , and the impact of this action on the group's future contribution (let us denote such a scheme by $S'(\gamma)$). Under this alternative scheme, the objectives of the participants remain aligned with that of the moderator, because the only component removed from the participant's objective function is not under his or her control, and free riding is reduced, because a participant does not get rewarded for the other participants' contribution unless it is tied to his or her own actions. This reduces the share of the created value that needs to be redistributed to the participants. (The impact

⁷ Note that I assume that the moderator's valuation for the ideas is high enough compared to c , such that it is possible and optimal for him or her to induce the participants to search in each period.

of participant i 's action in period t is defined here as: $\sum_{\tau=t+1}^{+\infty} \delta^{\tau-t} \lambda(t, \tau) \sum_{j \neq i} [C_{j\tau} - \tilde{C}_{j\tau}(i, t)]$ where

$\tilde{C}_{j\tau}(i, t)$ is the expected value of $C_{j\tau}$ obtained when y_{it} is null and all players maximize the group's expected output at each period, and $\lambda(t, \tau) = \delta^t \cdot (1 - \delta) / (\delta - \delta^\tau)$ is such that

$$\sum_{t=1}^{\tau-1} \lambda(t, \tau) = 1 \text{ for all } t \text{ and } \tau.$$

Proposition 5: $S'(1/2)$ is such that the objectives of the participants are aligned with that of the moderator, and the total expected payoffs distributed in the session are lower than under a proportional sharing rule.

Summary of this section

- Three stages can be defined in the idea generation process: exploration, exploitation and diversification.
- When the marginal contribution of new ideas following the same approach decreases quickly, exploration is never optimal and rewarding participants based on their individual contributions leads to optimal search.
- When the marginal contribution of new ideas following the same approach decreases slowly, exploration can be optimal and rewarding participants based on their individual contributions leads to suboptimal search.
- In the latter case, the group's output can be improved by rewarding each participant based on a weighted average between his or her individual contribution and his or her impact on the group's contribution.

This last result can be translated into a testable hypothesis. First, let us note that this result, like the others summarized above, assumes that all participants search for ideas at each period. Hence the different incentive schemes should be compared holding participation constant. The following hypothesis, which will be examined in the experiment in Section 4, is then sufficient for the result to hold:

Hypothesis: For a given level of participation, if the marginal contribution of new ideas following the same approach is slowly diminishing, the total contribution of the group is greater if participants are rewarded for their impact than it is if they are rewarded for their individual contributions.

The application and test of the insights from this theoretical analysis necessitate an idea generation system that is compatible with the incentive schemes considered in this section, i.e., which allows the measurements necessary to their implementation. Moreover, for the theoretical results to be managerially relevant, groups generating ideas using this system should perform better when incentives are present than when they are not, and perform better than groups or individuals generating ideas with other established methods. In particular, this idea generation system should not inhibit, and possibly enhance participants' creative abilities. Section 3 proposes a particular system.

3. An Ideation Game

I now describe an incentive-compatible "ideation game" that allows testing and implementing the insights derived in Section 2. In this game, participants score points for their ideas, the scoring scheme being adjustable to reward individual contribution, impact, or a weighted average of the two. The design of the game is based on the idea generation, bibliometric, and contract theory literatures. Table 1 provides a summary of the requirements imposed on the system and how they were addressed.

Table 1
Ideation game – requirements and corresponding solutions

Requirement	Proposed solution
Address "Production Blocking"	Asynchronous
Address "Fear of Evaluation"	Anonymous Objective measures Mutual monitoring
Measure contribution and impact	Ideas structured into trees
Prevent cheating	Relational contract Mutual monitoring

Addressing “fear of evaluation” and “production blocking”

Although the primary requirement imposed on this idea generation system is to allow implementing and testing the incentive schemes treated in Section 2, it should also be compatible with the existing literature on idea generation. In particular, two main issues have been identified with classical (face-to-face) idea generation sessions, in addition to free riding (Diehl and Stroebe 1987). The first one, “production blocking,” happens with classical idea generation sessions when participants are unable to express themselves simultaneously. The second one, “fear of evaluation,” corresponds to the fear of negative evaluation by the other participants, the moderator, or external judges. These two issues have been shown to be reduced by electronic idea generation sessions (Gallupe et al. 1991, Nunamaker et al. 1987), in which participation is asynchronous (therefore reducing production blocking) and in which the participants are anonymous (therefore reducing fear of evaluation). The online system proposed here is anonymous and asynchronous as well. In particular, participants create (or are given) an anonymous login and password, and log on to the idea generation session at their convenience (the session lasts typically for a few days).

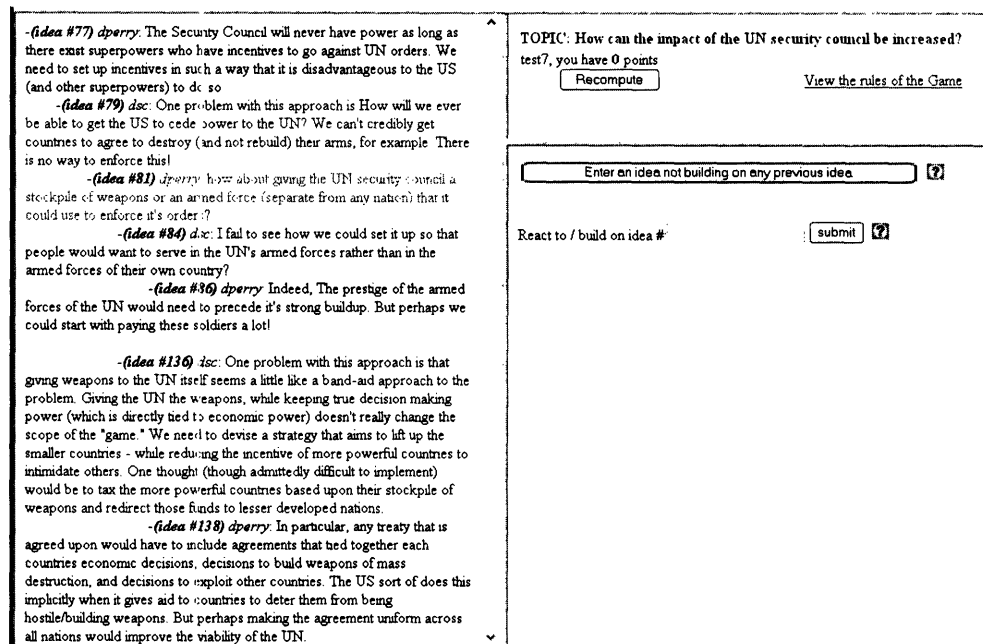
Measuring the contribution and the impact of participants

One simple way to measure the contribution and impact of participants would be to have them evaluated by external judges. However, this approach has limitations. First, it is likely to trigger fear of evaluation. Second, the evaluation criteria of the moderator, the judges and the participants might differ. Third, it increases the cost of the session. The proposed ideation game instead adopts a structure that provides objective measures of these quantities. These objective measures do not require additional evaluation, are computed based on well-stated rules, and offer continuous feedback on the performance of the participants.

Bibliometric research suggests that the number of ideas submitted by a participant should be a good measure of his or her contribution, and that his or her number of “citations” should be a good measure of his or her impact (King 1987). The proposed structure adapts the concept of citation count from the field of scientific publications (which are the focus of bibliometry), to idea generation. An example is provided in Figure 1. Ideas are organized into “trees,” such that the “son” of an idea appears below it, in a different color and with a different indentation (one level further to the right). The relation between a “father-idea” and a “son-idea” is that the son-

idea builds on its father-idea. More precisely, when a participant enters a new idea, he or she has the option to place it at the top of a new tree (by selecting “enter an idea not building on any previous idea”) or to build on an existing idea (by entering the identification number of the father-idea and clicking on “build on/react to idea #...”). Note that an idea can have more than one “son,” but at most one “father” (the investigation of more complex graph structures allowing multiple fathers is left for future research). The assignment of a father-idea to a new idea is left to the author of the new idea (see below how incentives are given to make the correct assignment).

Figure 1
Example of an ideation game



The impact of an idea y can then be measured by considering the ideas that are lower than y in the same tree. In the experiment described in the next section, impact was simply measured by counting these “descendant” ideas. More precisely, participants rewarded for their impact scored one point for each new idea that was entered in a tree below one of their ideas. For example, if idea #2 built on idea #1, and idea #3 on idea #2, then when idea #3 was posted, one point was given to the authors of both idea #1 and idea #2. Participants rewarded for their own contri-

butions scored one point per idea. The exploration of alternative measures is left for future research.

If a participant decides to build on an idea, a menu of “conjunctive phrases” (e.g. “More precisely”, “On the other hand”, “However”...) is offered to him or her (a blank option is also available), in order to facilitate the articulation of the links between ideas. Pretests suggested that these phrases were useful to the participants.

Preventing cheating

Although this structure provides potential measures of the contribution and impact of the different participants, if the scoring system were entirely automatic and if no subjective judgment were ever made, participants could “cheat” by simply posting irrelevant ideas in order to score unbounded numbers of points. Consequently, some level of subjective judgment seems inevitable. These judgments, however, should keep fear of evaluation to a minimum, and limit costly interventions by external judges. This is addressed by carefully defining the powers of the moderator of the session. First, the moderator does not have the right to decide which ideas deserve points. However, he or she has the power to expel participants who are found to cheat repeatedly, i.e., who submit ideas that clearly do not address the topic of the session or who wrongly assign new ideas to father-ideas. This type of relation between the participants and the moderator is called a “relational contract” in agency theory. In this relationship the source of motivation for the agent is the fear of termination of his or her relation with the principal (Baker et al. 1998, Levin 2003). In the present case, if the moderator checks the session often enough, it is optimal for participants to only submit ideas which they believe are relevant.

In order to reduce the frequency with which the moderator needs to check the session, this relational contract is complemented by a mutual monitoring mechanism (Knez and Simester 2001, Varian 1990). In particular, if participant i submits an idea which is found to be fraudulent by participant j , then j can “challenge” this idea.⁸ This freezes the corresponding stream of ideas (no participant can build on a challenged idea), until the moderator visits the session and determines who, between i and j , was right.⁹ The participant who was judged to be wrong then pays a fee to the other participant. The amount of this fee was fine-tuned based on pretests and was set

⁸ In the implementation used so far, participants cannot challenge an idea if it has already been built upon.

⁹ The idea is deleted if j was right.

to 5 points in the experiment. This mechanism is economical because it reduces the frequency with which the moderator needs to visit the session, without increasing its cost, since challenges result in a transfer of points between participants (no extra points need to be distributed). Beyond the cost factor, another argument in favor of having participants monitor each other is the finding by Collaros and Anderson (1969) that fear of evaluation is greater when the evaluation is done by an expert judge rather than by fellow participants. Note that in equilibrium, there should be no challenges, because participants should only submit relevant ideas and challenge irrelevant ideas. Indeed, in the experiment, there was only one challenge.

Practical implementation

This ideation game was programmed in php, using a MySQL database. Its structure and the instructions were fine-tuned based on two pretests. The pretests used participants from a professional marketing research/strategy firm. One pretest addressed an internal problem (improving the workspace) and a second pretest addressed an external problem (how to re-enter a market).

In the implementation, the moderator also has the ability to enter comments, either on specific ideas or on the overall session. Participants also have the ability to enter comments on specific ideas. This allows them, for example, to ask for clarification or to make observations that do not directly address the topic of the session. The difference between comments and ideas is that comments are not required to contribute positively to the session, and are not rewarded.

Finally, note that the implementation choices (e.g., amount of the fee in case of a challenge), although determined by the pretests, are likely to be improved with further experimentation.

4. Experiment

The experiment presented in this paper has two primary goals. The first goal is to test whether incentives have the power to improve the output of idea generation sessions. Recall that some research in social psychology suggests that incentives will have a negative influence on performance. The second goal is to verify the hypothesis, stated in Section 2, that under certain conditions, the total contribution of the group can be improved by rewarding each participant for the impact of his or her contribution, as opposed to his or her own contribution. This experiment also allows studying further the influence of incentives on the dynamics of idea generation.

Experimental design

The experiment used the ideation game described in the previous section, and had three conditions. The only difference between conditions was the incentive system, that is, the manner in which points were scored in the game. Each participant was randomly assigned to one idea generation session, and each idea generation session was assigned to one of the three conditions (all participants in a given session were given the same incentive scheme). The points scored during the session were later translated into cash rewards. Recall that the hypothesis derived in Section 2 relies on two assumptions: an equal level of participation, and a slowly diminishing marginal contribution. The latter is implied by the systematic measures of contribution and impact used in this experiment, which value all ideas equally. The former was addressed by calibrating the value of points in the different conditions based on a pretest, such that participants in all conditions could expect similar payoffs.¹⁰

A total of 78 participants took part in the experiment over a period of 10 days, signing up for the sessions at their convenience. Three sets of parallel sessions were run (one session per condition in each set), defining three “waves” of participants.¹¹ Each session lasted up to five days, and within each wave, the termination time of the three sessions was the same.¹²

The first condition (the “Flat” condition) was such that participants received a flat reward of \$10 for participation. In this condition, no points were scored (points were not even mentioned).

In the second condition (the “Own” condition), participants were rewarded based exclusively on their own contributions, and scored one point per idea submitted (each point was worth \$3).

¹⁰In this pretest, a group of approximately 15 employees from a market research firm generated ideas on “How can we better utilize our office space?”. Points were scored using the same scheme as in the “Impact” condition. Although points were not translated to monetary rewards in this pretest, participants found the game stimulating and felt directly concerned with the topic, resulting in 40 ideas submitted. Calibration was done conservatively, such that based on the pretest, participants in the “Flat” condition should expect slightly higher payoffs than participants in the “Own” condition, who in turn should expect slightly higher payoffs than participants in the “Impact” condition.

¹¹ The first wave consisted of the first 36 participants who signed up and had 12 participants per condition, the second wave consisted of the following 20 participants, with 7 participants in the “Flat” and in the “Own” conditions and 6 in the “Impact” condition, and the last wave consisted of the last 22 participants, with 8 in the “Flat” condition, and 7 in the “Own” and “Impact” conditions.

¹² This termination time was usually driven by budget considerations. Messages were posted on the “Impact” and “Own” sessions informing the participants that the session was over. The “Flat” sessions were not interrupted, but only the ideas submitted before the termination of the other sessions were taken into account.

In the third condition (the “Impact” condition), each participant was rewarded based exclusively on the impact of his or her ideas. Participants scored one point each time an idea was submitted that built on one of their own ideas (see previous section for the details of the scoring scheme). One interpretation of this scoring rule could be that participants scored one point for each “citation” of one of their ideas. Each point was worth \$2 for the first two waves of participants; the value of points was decreased by half to \$1 for the last wave. The results for the third wave were similar to those for the first two waves; hence the same analysis was applied to the data from the three waves.

In order to provide a strong test of the hypothesis that incentives can improve idea generation, an engaging topic, as well as some motivated participants, were selected. In particular, the topic was: “How can the impact of the UN Security Council be increased?” (note that the experiment was run in March 2003, at the time of the US-led war in Iraq), and the participants were recruited at an anti-war walkout in a major metropolitan area on the east coast, as well as on the campus of an east coast university.

Three graduate students in political science in the same university (naïve to the hypotheses) were later hired as expert judges, and were asked to evaluate independently the output of the sessions on several qualitative criteria (the judges were paid \$50 each for their work).

Verification of the theoretical hypothesis

As noted earlier, participants in this experiment were rewarded as if all ideas were equally valued by the moderator (there was no discounting of ideas depending on their position in a tree). More precisely, participants were rewarded as if the value of the parameter α from Section 2 was equal to 1, i.e., as if contribution was measured by the number of ideas.

With $\alpha=1$, the theoretical hypothesis proposed at the end of Section 2 suggests that the total number of ideas should be higher when participants are rewarded for their impact (measured by their number of “citations”), than when they are rewarded for their number of ideas. Note the counter-intuitive nature of this prediction: in the “Own” condition, although participants are trying to maximize their number of ideas and know that the other participants in the group have the same objective, they produce fewer ideas than the participants in the “Impact” condition who are rewarded for their impact on the group. The intuition behind this prediction is that participants

who are rewarded for their number of ideas are less likely to explore new approaches, which leads to less inspiring ideas, which leads to fewer ideas in total.

The data are consistent with this prediction: compared to the “Own” condition, participants in the “Impact” condition submitted significantly more unique ideas (p-value<0.05),¹³ where the number of unique ideas is defined as the number of ideas minus the number of redundant ideas (an idea is classified as redundant if it was judged as redundant by any of the three judges).

Other quantitative results

Both the quantitative and qualitative results are summarized in Table 2. The quantitative results are averaged across participants.

Compared to the “Own” condition, participants in the “Impact” condition not only submitted significantly more unique ideas, they were also significantly more likely to submit at least one unique idea (p-value<0.04), submitted significantly more unique ideas conditioning on submitting at least one (p-value<0.02), and submitted ideas that were on average 76% longer. Participants in the “Own” condition, in turn, performed better than participants in the “Flat” condition, although not significantly so on all metrics.¹⁴

Quality ratings

The qualitative results reported in Table 2 were obtained by averaging the ratings of the three judges.¹⁵ The ratings were found to have a reliability of 0.75¹⁶ (Rust and Cooil 1994), which is above the 0.7 benchmark often used in market research (Boulding et al. 1993).

¹³ One cannot assume that the number of ideas submitted by different participants are independent, since participants are influenced by the ideas submitted by the other members of their group. An endogeneity issue arises, because the number of ideas submitted by participant i depends on the number of ideas submitted by participant j , and vice versa. In order to limit this issue, a regression was run with the number of ideas submitted by participant i as the dependent variable, and three dummies (one per condition) as well as the number of ideas that were submitted by the other members of participant i 's group *before* he signed up as independent variables. This last independent variable is still endogenous, but it is pre-determined, leading to consistent estimates of the parameters (Greene 2000).

¹⁴ p-values below 0.18, 0.17, and 0.041 respectively for the number of unique ideas, for the proportion of participants who submitted at least one unique idea, and for the number of unique ideas given that at least one was submitted.

¹⁵ All ratings, except for the number of star ideas, are on a 10-point scale.

¹⁶ For each measure and for each judge, the ratings were normalized to range between 0 and 1. Then ratings were classified into three intervals: $[0;1/3[$, $[1/3;2/3[$, and $[2/3;1]$.

The sessions in the “Impact” condition were judged to have a larger overall contribution, more “star” ideas (i.e., very good ideas), more breadth, depth, and to have ideas that on average were more novel, thought-provoking and interactive (interactivity is defined here as the degree to which a “son-idea” builds on its “father-idea” and improves upon it).

Hence the results suggest that incentives do have the capacity to improve idea generation, and are consistent with the hypothesis formulated in Section 2.

Table 2
Results

	Impact	Own	Flat
Quantitative results			
Number of unique ideas per participant	4.6	2.3	0.8
Proportion of participants who posted at least one unique idea	68%	54%	48%
Number of unique ideas given that at least one	6.8	4.3	1.6
Number of words per idea	79.3	44.9	45.6
Qualitative ratings			
Total contribution	5.8	4.7	3.6
Number of star ideas	6.2	3.3	1.7
Breadth	6.1	5.3	3.2
Depth	6.8	4.6	3.6
Novelty	5.6	5.2	4.4
Thought-provoking	5.8	5.2	4.0
Interactivity	7.3	4.7	3.2

* = Impact significantly larger than “Own” as 0.05 level. † = Own significantly larger than Flat at 0.05 level.

Reconciling the results with the social psychology literature

As was mentioned in the introduction, the social psychology literature seems to predict that incentives are more likely to hurt idea generation. This prediction relies in great part on the observation that idea generation is not a task in which there exist easy algorithmic solutions, achievable by the straightforward application of certain operations. McGraw (1978) contrasts tasks such that the path to a solution is “well mapped and straightforward” with tasks such that

this path is more complicated and obscure. He argues that incentives are likely not to help with this second type of tasks, but notes that “it is nonetheless possible for reward to facilitate performance on such problems in the case where the algorithm (leading to a solution) is made obvious” (p. 54). Perhaps such facilitation might have happened in the present experiment. In particular, the structure of the ideation game might have made the mental steps leading to new ideas more transparent, allowing the participants to approach the task in a more “algorithmic” manner.

Incentives and motivation

Table 2 indicates that fewer participants in the “Flat” condition submitted at least one idea. Further analysis suggests that this might be due to a lower level of motivation of these participants. More precisely, 100% of the participants who submitted at least one unique idea in the “Flat” condition did so in the first 30 minutes after signing up, versus 65% for the “Impact” condition and 64% for the “Own” condition. In particular, 18% (respectively 21%) of the participants who submitted at least one idea in the “Impact” condition (respectively the “Own” condition) did so more than three hours after signing up. This is consistent with the hypothesis that all three conditions faced the same distribution of participants, but that when incentives were present, participants tried harder to generate ideas and did not give up as easily.

Dynamics of the sessions

The study of the dynamics of the ideation sessions, although not directly testing the theoretical predictions from Section 2, provides additional insight on the influence of incentives on the idea generation process. These exploratory analyses also provide direction for future research.

First, I study whether the quantitative results are due to participants posting more ideas at each visit, or visiting the session more often. Both the “Impact” and “Own” conditions give clear incentives to submit more ideas at each visit. In addition, we might expect participants in the “Impact” condition to have been more concerned with the evolution of the session, leading to more frequent visits. In the experiment, the time at which each idea was submitted was recorded. Let us then define “pockets” of ideas such that two successive ideas by the same participants are in the same pocket if they were submitted less than 20 minutes apart.¹⁷ Table 3 reports the aver-

¹⁷ Similar results were obtained when defining pockets using 10 or 30 minutes.

age number of ideas per pocket for the participants who posted at least one idea, as well as the average number of pockets. Consistent with our predictions, participants in the “Own” and “Impact” conditions submitted significantly more ideas at each of their visits to the session compared to participants in the “Flat” condition. Furthermore, participants in the “Impact” condition visited the session significantly more often.

Table 3
Dynamics

	Impact	Own	Flat
Number of ideas per pocket	2.9**	2.3*	1.2
Number of pockets per participant	2.6**	1.6	1.2

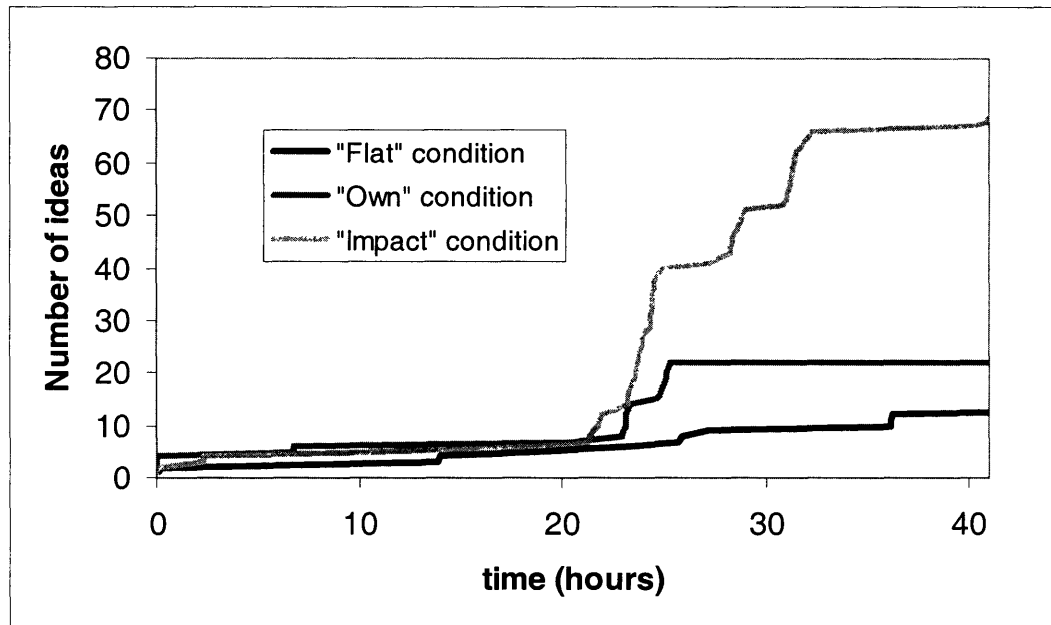
*significantly larger than “Flat,” $p < 0.1$
 **: significantly larger than “Flat,” $p < 0.03$

Next, I study how incentives influenced the overall number of ideas submitted in the session as a function of time. Figure 2 (corresponding to the first wave of participants - the graphs for the other two waves having similar characteristics) shows that the process in the “Impact” condition is S-shaped, whereas the process in the “Flat” condition tends to be smoother.¹⁸ The process in the “Own” condition is somewhat intermediary between these two extremes.

In order to better understand the shapes of these graphs, it is useful to plot similar graphs at the individual level. These graphs suggest that the differences in shape of the aggregate graphs reflects the fact that participants in the “Impact” condition relied more heavily on group interactions. In the “Impact” condition, the individual processes have the same characteristics as the aggregate graphs (period of inactivity followed by a period of high activity which finally stabilizes). The periods of high activity start at similar times for the different participants of a session, which accounts for the shape of the aggregate graphs. On the other hand, the groups in the “Flat” condition displayed no particular structure: participation is spread throughout the session and participants appear to be largely independent, resulting in smooth aggregate graphs.

¹⁸ Time on the x axis is the total cumulative time spent by participants in the session when the idea was submitted, divided by the number of participants. For example, if an idea was posted at 5 pm, and at this time two participants had signed up, one at 2 pm and one at 3 pm, then the reported time is $[(5-2)+(5-3)]/2=2.5$ hours. A simpler measure of time, the total clock time elapsed since the first participant signed up, gives similar qualitative results.

Figure 2
Number of ideas as a function of time



In conclusion, the study of the dynamics of the sessions indicates that incentives do not only influence the quantity and the quality of the ideas submitted, but also the manner in which they are submitted. This is reflected by both within- and between-participant patterns. The theoretical study of these patterns could be addressed in future research, for example by extending the model proposed in Section 2.

Limitations and alternative explanations

The hypothesis from Section 2 assumes that participation in the “Own” and “Impact” conditions was similar, i.e., participants spent comparable amounts of time searching for ideas.

If this assumption were not true, then this variation might provide an alternative explanation for the results. Three effects, potentially testable with future experiments, might lead to such alternative explanations: a non-monetary effect, an indirect incentives effect, and a direct incentives effect. A non-monetary effect might attribute higher participation in the “Impact” condition to a greater level of stimulation, resulting from richer interactions between participants. The indirect incentives effect is based on the hypothesis that, if ideas in the “Impact” condition were more inspiring, participants would be able to generate more ideas in the same amount of time,

making participation more attractive. The direct incentives effect hypothesizes that the expected payoff per idea was perceived as higher in the “Impact” condition. In particular, although the monetary value of points was calibrated to equate the expected payoff per idea in the two conditions, the calibration was based on a pretest using a different topic, and did not take into account participants’ subjective beliefs.

The non-monetary and indirect incentive effects would also advocate the use of impact as an incentive, but for different theoretical reasons. Further experiments could identify their magnitudes, providing direction for future theoretical research. For example, one might control the amount of time spent by the participants on the task by running the sessions in a laboratory. The third effect would make the comparison of the “Impact” and “Own” conditions problematic and cast some doubt on the validity of rewarding participants for their impact. In order to address this concern, another experiment could be run, in which the number of points earned by an idea in the “Impact” condition is bounded. For example, if ideas in this condition score points only for the first five subsequent ideas in the tree, and if the monetary value of a point is five times higher in the “Own” condition, then the expected payoff per idea is objectively lower in the “Impact” condition than it is in the “Own” condition.

5. Summary and Future Research

In this paper, I propose a quantitative framework for the study of a mostly qualitative topic, idea generation, with a focus on the effect of incentives. I first derive an analytical model of the idea generation process. Based on this model, I identify conditions under which rewarding participants for their individual contributions induces desirable actions, and other conditions under which it does not. In the latter case, performance can be improved by rewarding each participant for the impact of his or her contribution. I then develop an incentive compatible “ideation game,” which allows implementing the insights from the theoretical analysis and testing some of the predictions. Finally, I show experimentally that incentives have the capability to improve idea generation, in a manner consistent with the theory.

Several opportunities for future research can be identified in addition to the ones mentioned throughout the paper. First, it might be interesting to extend the theoretical model and take into account other dimensions of idea generation than the choice of which approach to follow. Second, it might be useful to generalize the experimental findings using different topics, and

other (potentially subjective) measures of the contribution and impact of the participants. Third, although the experimental findings in this paper are consistent with the theoretical prediction, they do not test the causal relationship between the speed of diminishing returns and the validity of rewarding impact, which could be addressed by additional experiments.

Finally, it would be interesting to identify conditions under which incentives are more likely to enhance or inhibit idea generation, as well as conditions under which the ideation game presented here is more or less likely to enhance participants' creativity. More generally, future research might examine how the experimental findings reported in this paper relate to classic results in social psychology. In a different paper (*author 2003*), I attempt to address these questions using major recent work in idea generation (Goldenberg et al. 1999a, Goldenberg et al. 1999b, Goldenberg and Mazursky 2002). In particular, there exist at least two opposing views regarding which cognitive skills should be encouraged in ideation sessions. The first view is that participants should be induced to think in a random fashion. This widely held belief, based on the assumption that anarchy of thought increases the probability of creative ideas, has led to the development of ideation tools such as brainstorming (Osborn 1957), Synectics (Prince 1970), and lateral thinking (De Bono 1970). In contrast, recent papers suggest that structure, and not randomness, is the key to creativity (Goldenberg et al. 1999a). This structured view, according to which creativity can be achieved through the identification and application of some regularities in previous creative output, has led to systematic approaches to idea generation. Two important examples are inventive templates (Goldenberg et al. 1999b, Goldenberg and Mazursky 2002), and TRIZ (Altshuller 2000). In *author 2003*, I try to shed some light on the apparent contradiction between these two views, and to identify conditions under which each one of them is more likely to be valid. I then show how this analysis applies to the questions mentioned above.

References

- Altshuller, Genrich (2000), *The Innovation Algorithm*, Technical Innovation Center, Inc., Worcester, MA.
- Baker, G., R. Gibbons and K. Murphy (1998), "Relational Contracts and the Theory of the Firm", mimeo, MIT.
- Bellman, R. (1961), *Adaptive Control Process: A guided Tour*, Princeton University Press.
- Boulding, William, Richard Staelin, Valarie Zeithaml, and Ajay Kalra (1993), "A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions", *Journal of Marketing Research*, 30 (February), 7-27.
- Collaros, Panayiota A., and Lynn R. Anderson (1969), "Effect of Perceived Expertness Upon Creativity Of Members of Brainstorming Groups", *Journal of Applied Psychology*, Vol. 53, No. 2, 159-163
- Dahan, Ely, and John R. Hauser (2001), "Product Development – Managing a Dispersed Process", in the *Handbook of Marketing*, Barton Weitz and Robin Wensley, Editors.
- De Bono, Edward (1970), *Lateral thinking: a textbook of creativity*, Ward Lock Educational, London.
- Diehl, Michael, and Wolfgang Stroebe (1987), "Productivity Loss in Brainstorming Groups: Toward the Solution of a Riddle", *Journal of Personality and Applied Psychology*, Vol. 53, No.3, 497-509
- Gallupe, Brent R., Lana M. Bastianutti, and William H. Cooper (1991), "Unblocking Brainstorms", *Journal of Applied Psychology*, vol. 76, No. 1, 137-142
- Gelman, Andrew B., John S. Carlin, Hal S. Stern, and Donald B. Rubin (1995), *Bayesian Data Analysis*, Chapman & Hall/CRC
- Goldenberg, J., D. Mazursky, and S.Solomon (1999a), "Creative Sparks", *Science*, 285 1495-1496.
- _____, _____, _____ (1999b), "Toward identifying the inventive templates of new products: A channeled ideation approach, *Journal of Marketing Research*, 36 (May) 200-210.
- _____, _____ (2002), *Creativity in product innovation*, Cambridge University Press.
- Greene, William H. (2000), *Econometric Analysis*, Prentice Hall International, Inc.

- Harkins, Stephen G., and Richard E. Petty (1982), "Effects of Task Difficulty and Task Uniqueness on Social Loafing", *Journal of Personality and Applied Psychology*, Vol. 43, No.6, 1214-1229
- Holmstrom, Bengt (1982), "Moral hazard in teams", *Bell Journal of Economics*, 13(2): 324-340
- Kerr, Norbert L., and Steven E. Bruun (1983), "Dispensability of Member Effort and Group Motivation Losses: Free-Rider Effects", *Journal of Personality and Applied Psychology*, Vol. 44, No.1, 78-94
- King, Jean (1987), "A Review of Bibliometric and Other Science Indicators and Their Role in Research Evaluation", *Journal of Information Science*, 13, 261-276
- Kuhn, Thomas S. (1977), "The Essential Tension: Tradition and Innovation in Scientific Research", In T. S. Kuhn (Ed.), *The essential tension: Selected readings in scientific tradition and change*, Chicago, IL.: University of Chicago Press.
- Knez, Marc, and Duncan Simester (2001), "Firm-Wide Incentives and Mutual Monitoring at Continental Airlines", *Journal of Labor Economics*, vol. 19, no. 4, 743-772
- Lamm, Helmut and Gisela Trommsdorff (1973), "Group versus individual performance on tasks requiring ideational proficiency (brainstorming): A review", *European Journal of Social Psychology*, 3 (4), 361-388
- Levin, J. (2003), "Relational Incentive Contracts", *American Economic Review*, forthcoming,
- McCullers, J.C. (1978), "Issues in learning and motivation", in M.R. Lepper & D. Greene (Eds.), *The hidden costs of reward* (pp. 5-18), Hillsdale, NJ: Erlbaum
- McGraw, K.O. (1978), "The detrimental effects of reward on performance: A literature review and a prediction model", in M.R. Lepper & D. Greene (Eds.), *The hidden costs of reward* (pp. 33-60), Hillsdale, NJ: Erlbaum
- Nunamaker, Jay F., Jr., Lynda M. Applegate, and Benn R. Konsynski (1987), "Facilitating group creativity: Experience with a group decision support system", *Journal of Management Information Systems*, Vol. 3 No. 4 (Spring).
- Osborn, A.F. (1957), *Applied Imagination*, (Rev. Ed.), New York: Scribner
- Prince, George M. (1970), *The Practice of Creativity; a manual for dynamic group problem solving*, New York, Harper & Row.
- Rust, Roland, and Bruce Cooil (1994), "Reliability Measures for Qualitative Data: Theory and Implications", *Journal of Marketing Research*, vol. XXXI (February 1994), 1-14.

- Simonton, Dean K. (1988), *Scientific Genius: A Psychology of Science*, Cambridge University Press
- Spence, K.W. (1956), *Behavior theory and conditioning*, New Haven: Yale University Press.
- Toubia, Olivier (2003), "Bounded rationality models of idea generation", working paper.
- Varian, Hal R. (1990), "Monitoring Agents with Other Agents", *Journal of Institutional and Theoretical Economics (JITE)*, Vol. 146, 153-174
- Von Hippel, Eric, and Ralph Katz (2002), "Shifting Innovation to Users via Toolkits", *Management Science*, vol. 48, No.7, pp 821-833.
- Williams, Kipling, Stephen Harkins, and Bibb Latané (1981), "Identifiability as a Deterrent to Social Loafing: Two Cheering Experiments", *Journal of Personality and Applied Psychology*, Vol. 40, No.2, 303-311
- Zajonc, Robert B. (1965), "Social Facilitation", *Science*, Vol. 149, No. 16 (July), 269-274

Appendix : proofs of the propositions

Let us define (we define similar quantities for B): $ST(A) = \alpha^{\hat{p}_A N_A} \cdot \hat{p}_A$ the expected payoff from period 1 obtained by choosing approach A in period 1; $S_A = \alpha^{\hat{p}_A N_A + 1} \cdot (\hat{p}_A N_A + 1) / (N_A + 1)$ the expected payoff from period 2 obtained by playing A in period 2 given that A was played in period 1 and was successful; $F_A = \alpha^{\hat{p}_A N_A} \cdot \hat{p}_A N_A / (N_A + 1)$ the expected payoff from period 2 obtained by playing A in period 2 given that A was played in period 1 and was unsuccessful; $LT(A) = \hat{p}_A \cdot \max\{S_A, ST(B)\} + (1 - \hat{p}_A) \cdot \max\{F_A, ST(B)\}$ the expected payoff from period 2, calculated at the beginning of period 1, if A is played in period 1.

Proof of Proposition 1a:

First, note that N_A and N_B are restricted to integer values, and that $\hat{p}_A \in [1/N_A, (N_A - 1)/N_A]$ and $\hat{p}_B \in [1/N_B, (N_B - 1)/N_B]$. Let us first show that for all \hat{p}_B, N_A , for all $\alpha < 1$, for all \hat{p}_A close enough to \hat{p}_B such that $\hat{p}_A < \hat{p}_B$, $ST(B) \leq ST(A) \Rightarrow N_B \geq N_A$. $ST(B)$ is decreasing in N_B . When $\hat{p}_A = \hat{p}_B$, $ST(B) > ST(A)$ if $N_B = N_A - 1$. By continuity, this is also true for \hat{p}_A close enough to \hat{p}_B such that $\hat{p}_A < \hat{p}_B$, hence $ST(B) \leq ST(A) \Rightarrow N_B \geq N_A$.

Then we have $ST(B) \leq ST(A) \Rightarrow S_B \leq S_A$. To see this, note that $S_B \leq S_A \Leftrightarrow \alpha^{\hat{p}_B N_B} \cdot (\hat{p}_B N_B + 1) / (N_B + 1) \leq \alpha^{\hat{p}_A N_A} \cdot (\hat{p}_A N_A + 1) / (N_A + 1) \Leftrightarrow \alpha^{\hat{p}_B N_B} \cdot \hat{p}_B + \alpha^{\hat{p}_B N_B} \cdot (1 - \hat{p}_B) / (N_B + 1) \leq \alpha^{\hat{p}_A N_A} \cdot \hat{p}_A + \alpha^{\hat{p}_A N_A} \cdot (1 - \hat{p}_A) / (N_A + 1) \Leftrightarrow \alpha^{\hat{p}_B N_B} \cdot (1 - \hat{p}_B) / (N_B + 1) \leq \alpha^{\hat{p}_A N_A} \cdot (1 - \hat{p}_A) / (N_A + 1)$ and $ST(B) \leq ST(A) \Leftrightarrow ST(B) \leq ST(A)$, $\alpha^{\hat{p}_B N_B} \leq \alpha^{\hat{p}_A N_A}$, $1 - \hat{p}_B < 1 - \hat{p}_A$, $N_B \geq N_A \Leftrightarrow ST(B) \leq ST(A)$, $\hat{p}_B > \hat{p}_A$.

Then, when $ST(B) \leq ST(A)$, there are 2 cases:

- $S_A \leq ST(B)$. This implies $S_B \leq S_A \leq ST(B) \leq ST(A)$. Then $LT(A) \geq ST(B)$ and $LT(B) = ST(A)$ and so $ST(B) + LT(B) = ST(B) + ST(A) \leq ST(A) + LT(A)$.

- $S_A > ST(B)$. Then $LT(A) \geq \hat{p}_A \cdot S_A + (1 - \hat{p}_A) \cdot ST(B)$ and $LT(B) = \hat{p}_B \cdot \max\{S_B, ST(A)\} + (1 - \hat{p}_B) \cdot ST(A)$

(because $F_B < ST(B) \leq ST(A)$). Then $ST(A) + LT(A) \geq ST(B) + LT(B)$ if $ST(A) + \hat{p}_A \cdot S_A + (1 - \hat{p}_A) \cdot ST(B) \geq ST(B) + \hat{p}_B \cdot \max\{S_B, ST(A)\} + (1 - \hat{p}_B) \cdot ST(A) \Leftrightarrow \hat{p}_A \cdot (S_A - ST(B)) \geq \hat{p}_B \cdot (\max\{S_B, ST(A)\} - ST(A))$

$$\begin{aligned}
ST(A) &\Leftrightarrow \hat{p}_A (S_A - ST(B)) \geq \hat{p}_B \cdot (S_B - ST(A)) \text{ (because } S_A > ST(B)) \Leftrightarrow \\
&\hat{p}_A \cdot [\alpha^{\hat{p}_A N_A + 1} \cdot (\hat{p}_A N_A + 1) / (N_A + 1) - \alpha^{\hat{p}_B N_B} \cdot \hat{p}_B] \geq \hat{p}_B \cdot [\alpha^{\hat{p}_B N_B + 1} \cdot (\hat{p}_B N_B + 1) / (N_B + 1)] - \\
&\alpha^{\hat{p}_A N_A} \cdot \hat{p}_A \Leftrightarrow \alpha^{\hat{p}_A N_A} \cdot \hat{p}_A \cdot [\alpha \cdot (\hat{p}_A N_A + 1) / (N_A + 1) + \hat{p}_B] \geq \alpha^{\hat{p}_B N_B} \cdot \hat{p}_B \cdot [\alpha \cdot (\hat{p}_B N_B + 1) / (N_B + 1) + \hat{p}_A] \Leftrightarrow \\
&\alpha \cdot (\hat{p}_A N_A + 1) / (N_A + 1) + \hat{p}_B \geq \alpha \cdot (\hat{p}_B N_B + 1) / (N_B + 1) + \hat{p}_A \text{ (because } ST(B) \leq ST(A)) \Leftrightarrow \\
&\hat{p}_B - \hat{p}_A \geq \alpha \cdot [(\hat{p}_B N_B + 1) / (N_B + 1) - (\hat{p}_A N_A + 1) / (N_A + 1)] \Leftrightarrow (\hat{p}_B N_B + 1) / (N_B + 1) \leq \\
&(\hat{p}_A N_A + 1) / (N_A + 1) \text{ or } \{ (\hat{p}_B N_B + 1) / (N_B + 1) > (\hat{p}_A N_A + 1) / (N_A + 1) \text{ and } (\hat{p}_A N_A + 1) / (N_A + 1) - \\
&\hat{p}_A \geq (\hat{p}_B N_B + 1) / (N_B + 1) - \hat{p}_B \}, \text{ which is satisfied because } (\hat{p}_A N_A + 1) / (N_A + 1) - \\
&\hat{p}_A \geq (\hat{p}_B N_B + 1) / (N_B + 1) - \hat{p}_B \Leftrightarrow (1 - \hat{p}_A) / (N_A + 1) \geq (1 - \hat{p}_B) / (N_B + 1) \Leftrightarrow \hat{p}_B > \hat{p}_A \text{ and } N_B \geq N_A.
\end{aligned}$$

Proof of Proposition 1b:

For all $\hat{p}_B, \hat{p}_A < \hat{p}_B, N_A$, let $N_B' = \min\{N_B: \hat{p}_B \cdot N_B > \hat{p}_A \cdot N_A\}$. If α is small enough, then $N_B \geq N_B' \Rightarrow ST(A) > ST(B)$. Indeed, $ST(A) > ST(B) \Leftrightarrow \alpha^{\hat{p}_A N_A} \cdot \hat{p}_A > \alpha^{\hat{p}_B N_B} \cdot \hat{p}_B \Leftrightarrow \alpha^{\hat{p}_B N_B - \hat{p}_A N_A} < \hat{p}_A / \hat{p}_B \Leftrightarrow \alpha^{\hat{p}_B N_B - \hat{p}_A N_A} < 1/N_A$ (because $\hat{p}_A \geq 1/N_A$ and $\hat{p}_B < 1$) $\Leftrightarrow \alpha < \alpha'$ where α' is such that $\alpha'^{\hat{p}_B N_B - \hat{p}_A N_A} = 1/N_A$. Then $ST(B) \geq ST(A) \Rightarrow N_B < N_B' \Rightarrow \hat{p}_B \cdot N_B \leq \hat{p}_A \cdot N_A$. Then if α is small enough, we have $S_A < ST(B)$ and $F_A < ST(A) \leq ST(B)$, which implies $LT(A) = ST(B)$. Then since $ST(A) \leq LT(B)$, we have $ST(A) + LT(A) \leq ST(B) + LT(B)$.

Proof of Proposition 1c:

Given \hat{p}_B, N_B must be such that $1 \leq \hat{p}_B \cdot N_B \leq N_B - 1$, i.e., N_B must be at least as large as $\max\{1/\hat{p}_B, 1/(1 - \hat{p}_B)\} = N_B^{min}$. Let us show that for all \hat{p}_B , for all N_A , for all α close enough to 1 such that $\alpha < 1$, for all \hat{p}_A close enough to \hat{p}_B such that $\hat{p}_A < \hat{p}_B$, we have the following:

1. $ST(B) > ST(A)$ for $N_B = N_B^{min}$ and $\lim_{N_B \rightarrow +\infty} ST(B) < ST(A)$
2. $\lim_{N_B \rightarrow +\infty} LT(B) < LT(A)$
3. $ST(B)$ is monotonically decreasing in N_B
4. $LT(B)$ is monotonically non-increasing in N_B
5. $ST(B) = ST(A)$ implies $LT(B) < LT(A)$

These conditions imply that there exists $N_B^{min} \leq N_B^l < N_B^{**}$ such that $N_B > N_B^{**} \Leftrightarrow ST(B) < ST(A), ST(B) + LT(B) - [ST(A) + LT(A)]$ is decreasing in N_B , negative for $N_B > N_B^{**}$, positive for $N_B < N_B^l$, and so by continuity and monotonicity there exists N_B^* such that $N_B^* < N_B^{**}$ and such that $ST(B) + LT(B) - [ST(A) + LT(A)]$ is non-negative for $N_B \leq N_B^*$ (exploitation), non-positive for $N_B^* \leq N_B \leq N_B^{**}$ (exploration) and negative for $N_B > N_B^{**}$ (diversification). Note that if $ST(B) + LT(B) < ST(A) + LT(A)$ for $N_B = N_B^{min}$, then $N_B^* < N_B^{min}$.

Let us prove conditions 1 to 5:

1) for $N_B = N_B^{min}$, $ST(B) = \alpha^{\hat{p}_B \cdot N_B^{min}} \hat{p}_B > ST(A) = \alpha^{\hat{p}_A N_A} \hat{p}_A$ for all $\alpha < 1$ and \hat{p}_A close enough to \hat{p}_B , and $\lim_{N_B \rightarrow \infty} ST(B) = 0 < ST(A)$.

2) When N_B goes to $+\infty$, the limits are: $LT(B) = ST(A) = \alpha^{\hat{p}_A N_A} \cdot \hat{p}_A$, $LT(A) = \hat{p}_A \cdot S_A + (1 - \hat{p}_A) \cdot F_A = \alpha^{\hat{p}_A N_A} \cdot \hat{p}_A \cdot (\hat{p}_A \cdot N_A \cdot (\alpha - 1) + N_A + \alpha) / (N_A + 1) < LT(B)$ because $\hat{p}_A \cdot N_A \cdot (\alpha - 1) + N_A + \alpha < N_A + 1$

3) is trivial

4) First let us note that S_B is decreasing in N_B because $N_B \rightarrow \alpha^{\hat{p}_B N_B + 1}$ is positive and decreasing in N_B , and so is $N_B \rightarrow (\hat{p}_B \cdot N_B + 1) / (N_B + 1)$ (the derivative of this last function with respect to N_B is $(\hat{p}_B - 1) / (N_B + 1)^2$). Next we have: $S_B < F_B \Leftrightarrow N_B > N_B^L = \alpha / ((1 - \alpha) \cdot \hat{p}_B)$. If $\alpha > \hat{p}_B \cdot N_A / (\hat{p}_B \cdot N_A + 1)$ then $N_A < N_B^L$, in which case $N_B > N_B^L \Rightarrow F_B = \alpha^{\hat{p}_B N_B} \cdot \hat{p}_B N_B / (N_B + 1) < \alpha^{\hat{p}_B N_B^L} \cdot \hat{p}_B < ST(A) = \alpha^{\hat{p}_A N_A} \cdot \hat{p}_A$ if \hat{p}_A close enough to \hat{p}_B .

Then we have 4 cases:

- $ST(A) > F_B > S_B$: $LT(B)$ is constant in N_B
- $ST(A) > S_B > F_B$: $LT(B)$ is constant in N_B
- $S_B > ST(A) > F_B$: $LT(B) = \hat{p}_B \cdot S_B + (1 - \hat{p}_B) \cdot ST(A)$ is decreasing in N_B because S_B is
- $S_B > F_B > ST(A)$: $LT(B) = \hat{p}_B \cdot S_B + (1 - \hat{p}_B) \cdot F_B = \alpha^{\hat{p}_B N_B} \cdot \hat{p}_B \cdot (N_B \cdot (\alpha \cdot \hat{p}_B + 1 - \hat{p}_B) + \alpha) / (N_B + 1)$ and $\partial LT(B) / \partial N_B =$

$$\frac{\hat{p}_B \cdot \ln(\alpha) \cdot \alpha^{\hat{p}_B N_B} \cdot (N_B + 1) - \alpha^{\hat{p}_B N_B}}{(N_B + 1)^2} \cdot \hat{p}_B \cdot (N_B \cdot (\alpha \cdot \hat{p}_B + 1 - \hat{p}_B) + \alpha) + \frac{\alpha^{\hat{p}_B N_B}}{N_B + 1} \cdot \hat{p}_B \cdot (\alpha \cdot \hat{p}_B + 1 - \hat{p}_B)$$

which is of the same sign as: $(\hat{p}_B \cdot \ln(\alpha) \cdot (N_B + 1) - 1) \cdot (N_B \cdot (\alpha \cdot \hat{p}_B + 1 - \hat{p}_B) + \alpha) / (N_B + 1) + (\alpha \cdot \hat{p}_B + 1 - \hat{p}_B)$ which is equal to 0 when $\alpha = 1$, and which is of the same sign, when α close to 1 (using a Taylor's series expansion), as $-\hat{p}_B \cdot N_B + (1 + \hat{p}_B N_B) / (N_B + 1) - 2 \cdot \hat{p}_B \leq$

$1 - 2 \cdot \hat{p}_B + N_B / (N_B + 1) < 0$. So $LT(B)$ is non-increasing in N_B if α close enough to 1 and \hat{p}_A close enough to \hat{p}_B .

5) The proof is similar to that of proposition 1: $ST(B)=ST(A) \Rightarrow N_B > N_A$ and $S_B < S_A$. Then if α close enough to 1, $S_A > ST(A)=ST(B)$ for all $\hat{p}_A \in [1/N_A, (N_A-1)/N_A]$, and $ST(A)+LT(A) > ST(B)+LT(B) \Leftrightarrow \hat{p}_B - \hat{p}_A > \alpha \cdot [(\hat{p}_B N_B + 1)/(N_B + 1) - (\hat{p}_A N_A + 1)/(N_A + 1)]$ which is satisfied for $\alpha < 1$ and $\hat{p}_B > \hat{p}_A$.

Proof of proposition 2:

Let us assume that if both players find an idea using the same approach in a given period, then we determine randomly which player submits his or her idea first. This is equivalent to having player's expected payoff equal to $(\alpha^n + \alpha^{n+1}) / (2p^2) + \alpha^n \cdot p \cdot (1-p)$ if n ideas have been submitted using the approach at the beginning of the period and if the probability of success is p . Recall that \hat{p}_A and \hat{p}_B are respectively in $[1/N_A, (N_A-1)/N_A]$ and $[1/N_B, (N_B-1)/N_B]$. Since the lower (resp. upper) bounds of the intervals above are strictly positive (resp. strictly smaller than 1), there exists α small enough such that in period 2 both players will play the approach with the smallest number of successes so far (no matter what the expected probabilities are, they cannot counter-balance the effect of the discount rate). In the special case in which the two approaches have the same number of successes, then players will choose different approaches if \hat{p}_A is close enough to \hat{p}_B (because choosing the same approach leads to much lower payoffs if α is small enough).

Let us first consider the case $N_A > N_B$. Then if \hat{p}_A is close enough to \hat{p}_B , $n_{SA} > n_{SB}$. Let us show that if α is small enough, then the only SPE is for both players to play B in period 1, and this is also socially optimal:

- $BR(B)=B: (\alpha^{n_{SB}} + \alpha^{n_{SB}+1}) / 2 \cdot E(p_B^2) + \alpha^{n_{SB}} \cdot E(p_B \cdot (1-p_B)) + LT(B, B) > \alpha^{n_{SA}} \cdot \hat{p}_A + LT(A, B)$

for α small enough if (if α is small enough then any term discounted by more than $\alpha^{n_{SB}}$ is negligible): $ST(B, B) + E((1-p_B)^2) \cdot ST(B, B | n_{FB} + 2) > (1 - \hat{p}_B) \cdot ST(B, B | n_{FB} + 1)$ where $ST(B, B) = (1 + \alpha) \cdot E(p_B^2) / 2 + E(p_B \cdot (1-p_B))$ is the expected instantaneous payoff divided by $\alpha^{n_{SB}}$ if players play (B, B) and $ST(B, B | n_{SF} + 1)$ is the similar quantity if both players play (B, B) after a failure in B . This inequality holds because $ST(B, B) + E[(1-p_B)^2] \cdot ST(B, B | n_{FB} + 2) > ST(B, B) > ST(B, B | n_{FB} + 1) > (1 - \hat{p}_B) \cdot ST(B, B | n_{FB} + 1)$

- BR(A)=B: as above, this is true for α small enough if $\hat{p}_B + (1 - \hat{p}_B) \cdot ST(B, B | n_{FB} + 1) > 0 + ST(B, B)$ which holds because $\hat{p}_B + (1 - \hat{p}_B) \cdot ST(B, B | n_{FB} + 1) > \hat{p}_B > ST(B, B)$
- (B,B) is socially optimal if α small enough because it is more likely that at least one idea will be found using B if both players play B in period 1.

The same argument applies to the case $N_A \leq N_B$, in which case $n_{SA} < n_{SB}$ if \hat{p}_A is close enough to \hat{p}_B .

Proof of proposition 3:

Let us consider the subgame perfect equilibrium of the game that consists in choosing which approach to play at $t=1$ and $t=2$.

In period 2, players each play the approach with the highest expected probability of success.

In period 1:

If player 2 plays A at $t=1$, then player 1 gets:

By playing A:

$$\hat{p}_A + E(p_A^2) \cdot \max\left\{\frac{n_{AS} + 2}{N_A + 2}, \hat{p}_B\right\} + 2E(p_A \cdot (1 - p_A)) \cdot \max\left\{\frac{n_{AS} + 1}{N_A + 2}, \hat{p}_B\right\} + E((1 - p_A)^2) \cdot \hat{p}_B$$

$$\text{By playing B: } \hat{p}_B + \hat{p}_A \cdot \hat{p}_B \cdot \max\left\{\frac{n_{AS} + 1}{N_A + 1}, \frac{n_{BS} + 1}{N_B + 1}\right\} + \hat{p}_A \cdot (1 - \hat{p}_B) \cdot \max\left\{\frac{n_{AS} + 1}{N_A + 1}, \frac{n_{BS} + 1}{N_B + 1}\right\}$$

$$+ (1 - \hat{p}_A) \cdot \hat{p}_B \cdot \frac{n_{BS} + 1}{N_B + 1} + (1 - \hat{p}_A) \cdot (1 - \hat{p}_B) \cdot \max\left\{\frac{n_{AS} + 1}{N_A + 1}, \frac{n_{BS} + 1}{N_B + 1}\right\}$$

If player 2 plays B at $t=1$, then player 1 gets:

$$\text{By playing A: } \hat{p}_A + \hat{p}_A \cdot \hat{p}_B \cdot \max\left\{\frac{n_{AS} + 1}{N_A + 1}, \frac{n_{BS} + 1}{N_B + 1}\right\} + \hat{p}_A \cdot (1 - \hat{p}_B) \cdot \max\left\{\frac{n_{AS} + 1}{N_A + 1}, \frac{n_{BS} + 1}{N_B + 1}\right\}$$

$$+ (1 - \hat{p}_A) \cdot \hat{p}_B \cdot \frac{n_{BS} + 1}{N_B + 1} + (1 - \hat{p}_A) \cdot (1 - \hat{p}_B) \cdot \max\left\{\frac{n_{AS} + 1}{N_A + 1}, \frac{n_{BS} + 1}{N_B + 1}\right\}$$

By playing B:

$$\hat{p}_B + E(p_B^2) \cdot \frac{n_{BS} + 2}{N_B + 2} + 2E(p_B \cdot (1 - p_B)) \cdot \max\left\{\hat{p}_A, \frac{n_{BS} + 1}{N_B + 2}\right\} + E((1 - p_B)^2) \cdot \max\left\{\hat{p}_A, \frac{n_{BS} + 1}{N_B + 2}\right\}$$

Let us consider the following condition: $\hat{p}_B < (n_{AS} + 1)/(N_A + 1)$ (a).

For all $\hat{p}_A < \hat{p}_B$ and N_A such that (a) is satisfied, if N_B is large enough (the beliefs on bandit B are stable enough), then we have that the following conditions are sufficient for both players to choose B at $t=1$ in any SPE of the game and for it not to be socially optimal:

- **BR(B)=B:** $\hat{p}_A + \hat{p}_A \cdot (n_{AS} + 1)/(N_A + 1) + (1 - \hat{p}_A) \cdot \hat{p}_B < \hat{p}_B + \hat{p}_B \Leftrightarrow \hat{p}_B - \hat{p}_A > \hat{p}_A \cdot [(n_{AS} + 1)/(N_A + 1) - \hat{p}_B]$ (b)
- **BR(A)=B:** $\hat{p}_B + (\hat{p}_A \cdot n_{AS} + 1)/(N_A + 1) + (1 - \hat{p}_A) \cdot \hat{p}_B > \hat{p}_A + E(p_A^2) \cdot (n_{AS} + 2)/(N_A + 2) + 2 \cdot E(p_A \cdot (1 - p_A)) \cdot \max\{\hat{p}_B, (n_{AS} + 1)/(N_A + 2)\} + E((1 - p_A)^2) \cdot \hat{p}_B$ (1)

If $\hat{p}_B > (n_{AS} + 1)/(N_A + 2)$ (c) then (1) becomes: $\hat{p}_B - \hat{p}_A > E(p_A^2) \cdot (\hat{p}_A \cdot N_A + 2)/(N_A + 2)$

$$- E(p_A^2) \cdot \hat{p}_B - \hat{p}_A \cdot (\hat{p}_A N_A + 1)/(N_A + 1) + \hat{p}_A \cdot \hat{p}_B \quad (\text{f})$$

If $\hat{p}_B < (n_{AS} + 1)/(N_A + 2)$ (c') then (1) becomes: $\hat{p}_B + \hat{p}_A \cdot (n_{AS} + 1)/(N_A + 1) + (1 - \hat{p}_A) \cdot \hat{p}_B$

$> \hat{p}_A + E(p_A^2) \cdot (n_{AS} + 2)/(N_A + 2) + 2E(p_A(1 - p_A)) \cdot (n_{AS} + 1)/(N_A + 2) + E((1 - p_A)^2) \cdot \hat{p}_B \Leftrightarrow \hat{p}_B - \hat{p}_A >$

$$E(p_A - p_A^2) \cdot \left(\frac{n_{AS} + 1}{N_A + 2} - \hat{p}_B\right) + \hat{p}_A \cdot \left(\frac{n_{AS} + 1}{N_A + 2} - \frac{n_{AS} + 1}{N_A + 1}\right) + E(p_A^2) \cdot \left(\frac{n_{AS} + 2}{N_A + 2} - \frac{n_{AS} + 1}{N_A + 2}\right)$$

The right-hand side is smaller than $\hat{p}_A \cdot [(n_{AS} + 1)/(N_A + 2) - \hat{p}_B]$ because $\hat{p}_A \cdot \left(\frac{n_{AS} + 1}{N_A + 2} - \frac{n_{AS} + 1}{N_A + 1}\right) +$

$$E(p_A^2) \cdot \left(\frac{n_{AS} + 2}{N_A + 2} - \frac{n_{AS} + 1}{N_A + 2}\right) = -\hat{p}_A \cdot \frac{n_{AS} + 1}{(N_A + 1) \cdot (N_A + 2)} + \frac{E(p_A^2)}{N_A + 2}$$

$$= \frac{1}{N_A + 2} \cdot (E(p_A^2) - \hat{p}_A \cdot \frac{\hat{p}_A \cdot N_A + 1}{N_A + 1}) = 0 \text{ because } E(p_A^2) = \hat{p}_A^2 + \hat{p}_A \cdot (1 - \hat{p}_A)/(N_A + 1) =$$

$\hat{p}_A \cdot (\hat{p}_A \cdot N_A + 1)/(N_A + 1)$ ($\text{Var}(p_A) = \hat{p}_A \cdot (1 - \hat{p}_A)/(N_A + 1)$). So the condition holds if

$$\hat{p}_A \cdot [(n_{AS} + 1)/(N_A + 2) - \hat{p}_B] < \hat{p}_B - \hat{p}_A \quad (\text{e})$$

- **A-B socially better than B-B:** $\hat{p}_A + \hat{p}_B + 2 \cdot \hat{p}_A \cdot (n_{AS} + 1)/(N_A + 1) + 2 \cdot (1 - \hat{p}_A) \cdot \hat{p}_B > 4 \cdot \hat{p}_B$
 $\Leftrightarrow \hat{p}_B - \hat{p}_A < 2 \cdot \hat{p}_A \cdot [(n_{AS} + 1)/(N_A + 1) - \hat{p}_B]$ (d)

• We'll also be using the following, when (c) is satisfied (A-A is socially better than B-B):

$$2 \cdot \hat{p}_A + 2E(p_A^2) \cdot (\hat{p}_A N_A + 2)/(N_A + 2) + 2 \cdot (1 - E(p_A^2)) \cdot \hat{p}_B > 4 \cdot \hat{p}_B \Leftrightarrow \hat{p}_B - \hat{p}_A <$$

$$\hat{p}_A \cdot [(\hat{p}_A N_A + 2)/(N_A + 2) - \hat{p}_B] \quad (\text{d}')$$

Let us consider the following conditions:

$$\hat{p}_B < (\hat{p}_A N_A + 1)/(N_A + 1) \Leftrightarrow N_A < (1 - \hat{p}_B)/(\hat{p}_B - \hat{p}_A) \quad (\text{a})$$

$$\hat{p}_A \cdot \left(\frac{\hat{p}_A \cdot N_A + 1}{N_A + 1} - \hat{p}_B \right) < \hat{p}_B - \hat{p}_A \Leftrightarrow N_A > \frac{2 - \hat{p}_B / \hat{p}_A - \hat{p}_B}{\hat{p}_B / \hat{p}_A - 1 + \hat{p}_B - \hat{p}_A} \quad (\text{b})$$

$$\hat{p}_B > (\hat{p}_A \cdot N_A + 1) / (N_A + 2) \Leftrightarrow N_A > (1 - 2 \cdot \hat{p}_B) / (\hat{p}_B - \hat{p}_A) \quad (\text{c})$$

$$N_A < (1 - 2 \cdot \hat{p}_B) / (\hat{p}_B - \hat{p}_A) \quad (\text{c}')$$

$$2 \hat{p}_A \cdot [(\hat{p}_A \cdot N_A + 1) / (N_A + 1) - \hat{p}_B] > \hat{p}_B - \hat{p}_A \Leftrightarrow N_A < \frac{3 - \hat{p}_B / \hat{p}_A - 2 \cdot \hat{p}_B}{\hat{p}_B / \hat{p}_A - 1 + 2 \cdot (\hat{p}_B - \hat{p}_A)} \quad (\text{d})$$

$$\hat{p}_B - \hat{p}_A < \hat{p}_A^2 \cdot [(\hat{p}_A \cdot N_A + 2) / (N_A + 2) - \hat{p}_B] \Leftrightarrow N_A < \frac{2 \cdot (1 - \hat{p}_B) \cdot \hat{p}_A^2 - 2 \cdot (\hat{p}_B - \hat{p}_A)}{(\hat{p}_B - \hat{p}_A) \cdot (1 + \hat{p}_A^2)} \quad (\text{d}')$$

$$\hat{p}_A \cdot [(\hat{p}_A \cdot N_A + 1) / (N_A + 2) - \hat{p}_B] < \hat{p}_B - \hat{p}_A \Leftrightarrow N_A > \frac{3 - 2 \hat{p}_B / \hat{p}_A - 2 \cdot \hat{p}_B}{\hat{p}_B / \hat{p}_A - 1 + (\hat{p}_B - \hat{p}_A)} \quad (\text{e})$$

$$\hat{p}_B - \hat{p}_A > E(p_A^2) \cdot \frac{\hat{p}_A \cdot N_A + 2}{N_A + 2} - E(p_A^2) \cdot \hat{p}_B - \hat{p}_A \cdot \frac{\hat{p}_A \cdot N_A + 1}{N_A + 1} + \hat{p}_A \cdot \hat{p}_B \quad (\text{f})$$

(a),(b),(c),(d'), (f) as well as (a),(b),(c'),(d),(e) are sets of sufficient conditions. Let us consider \hat{p}_A such that $(\hat{p}_B - \hat{p}_A) = \gamma(1 - \hat{p}_B)$ with γ small enough such that the numerators in (d) and (d') are positive.

Let us look at the first set of conditions (a),(b),(c),(d'), (f):

(f) imposes a lower bound on N_A .

$$\begin{aligned} (\text{f}) \Leftrightarrow & N_A^2 \cdot (\hat{p}_B - \hat{p}_A + E(p_A^2) \cdot \hat{p}_B - \hat{p}_A \cdot \hat{p}_B - E(p_A^2) \cdot \hat{p}_A + \hat{p}_A^2) + \\ & N_A \cdot [3 \cdot (\hat{p}_B - \hat{p}_A + E(p_A^2) \cdot \hat{p}_B - \hat{p}_A \cdot \hat{p}_B) - E(p_A^2) \cdot \hat{p}_A + \hat{p}_A - 2 \cdot E(p_A^2) + 2 \hat{p}_A^2] + \\ & 2 \cdot (\hat{p}_B + E(p_A^2) \cdot \hat{p}_B - \hat{p}_A \cdot \hat{p}_B - E(p_A^2)) > 0 \end{aligned}$$

Let $N_A^{\max} = \frac{2 \cdot \hat{p}_A^2 - 2 \cdot \gamma}{\gamma \cdot (1 + \hat{p}_A^2)}$ be the value of N_A defined by (d'). Then (f) and (d') can be satisfied

simultaneously when γ small enough if $(E(p_A^2) \rightarrow \hat{p}_B^2 \text{ when } \gamma \rightarrow 0)$:

$2 \cdot \hat{p}_B^2 \cdot (1 - \hat{p}_B) \cdot (1 - \hat{p}_B + \hat{p}_B^2) / (1 + \hat{p}_B^2) + \hat{p}_B \cdot (1 - \hat{p}_B) \cdot (1 - 2 \cdot \hat{p}_B) > 0 \Leftrightarrow 1 - \hat{p}_B^2 > 0$ which is satisfied for $\hat{p}_B < 1$. $1/\gamma$ is the value of N_A defined by (a). Then (f) and (a) can be satisfied simultaneously

when γ small enough if: $(1 - \hat{p}_B) \cdot (1 - \hat{p}_B + \hat{p}_B^2) + \hat{p}_B \cdot (1 - \hat{p}_B) \cdot (1 - 2 \cdot \hat{p}_B) > 0 \Leftrightarrow 1 - \hat{p}_B^2 > 0$ which is satisfied for $\hat{p}_B < 1$. (b) and (a) can be satisfied simultaneously if:

$$\frac{2 - \hat{p}_B / \hat{p}_A - \hat{p}_B}{\hat{p}_B / \hat{p}_A - 1 + \hat{p}_B - \hat{p}_A}$$

$\frac{1 - \hat{p}_B - (\hat{p}_B/\hat{p}_A - 1)}{\hat{p}_B - \hat{p}_A + (\hat{p}_B/\hat{p}_A - 1)} < (1 - \hat{p}_B)/(\hat{p}_B - \hat{p}_A)$ which is satisfied. (c) and (a) can be satisfied simul-

taneously because $(1 - 2\hat{p}_B)/(\hat{p}_B - \hat{p}_A) < (1 - \hat{p}_B)/(\hat{p}_B - \hat{p}_A)$. (b) and (d') can be satisfied simul-

taneously if $\frac{1 - \gamma/\hat{p}_A}{\gamma(1 + 1/\hat{p}_A)} < \frac{2\hat{p}_A^2 - 2\gamma}{\gamma(1 + \hat{p}_A^2)}$ which is satisfied for γ small enough if

$\frac{1}{1 + 1/\hat{p}_B} < \frac{2\hat{p}_B^2}{1 + \hat{p}_B^2} \Leftrightarrow \hat{p}_B^2 + 2\hat{p}_B - 1 > 0 \Leftrightarrow \hat{p}_B > -1 + \sqrt{2}$. (c) and (d') can be satisfied simultane-

ously if: $\frac{1 - 2\hat{p}_B}{\gamma(1 - \hat{p}_B)} < \frac{2\hat{p}_A^2 - 2\gamma}{\gamma(1 + \hat{p}_A^2)}$ which is satisfied for γ small enough if $(1 - 2\hat{p}_B)/(1 - \hat{p}_B) <$

$2\hat{p}_B^2/(1 + \hat{p}_B^2) \Leftrightarrow \hat{p}_B^2 + 2\hat{p}_B - 1 > 0 \Leftrightarrow \hat{p}_B > -1 + \sqrt{2}$. So (a),(b),(c),(d'), (f) can be simultaneously satisfied if $\hat{p}_B > -1 + \sqrt{2}$.

Now let us look at the other set of sufficient conditions (a),(b),(c'),(d),(e): first, (c') implies (a) so we can ignore (a). Then (e) and (c') can be satisfied simultaneously

if: $\frac{3 - 2\hat{p}_B/\hat{p}_A - 2\hat{p}_B}{\hat{p}_B/\hat{p}_A - 1 + (\hat{p}_B - \hat{p}_A)} = \frac{1 - 2\hat{p}_B - 2(\hat{p}_B/\hat{p}_A - 1)}{(\hat{p}_B - \hat{p}_A) + (\hat{p}_B/\hat{p}_A - 1)} < \frac{1 - 2\hat{p}_B}{\hat{p}_B - \hat{p}_A}$, which holds. (e) and (d) can be

satisfied simultaneously if: $\frac{3 - 2\hat{p}_B/\hat{p}_A - 2\hat{p}_B}{\hat{p}_B/\hat{p}_A - 1 + (\hat{p}_B - \hat{p}_A)} < \frac{3 - \hat{p}_B/\hat{p}_A - 2\hat{p}_B}{\hat{p}_B/\hat{p}_A - 1 + 2(\hat{p}_B - \hat{p}_A)}$

$\Leftrightarrow \frac{(1 - \hat{p}_B)(2 - 2\gamma/\hat{p}_A) - 1}{(1 - \hat{p}_B)\gamma(1 + 1/\hat{p}_A)} < \frac{(1 - \hat{p}_B)(2 - \gamma/\hat{p}_A)}{(1 - \hat{p}_B)\gamma(2 + 1/\hat{p}_A)}$ which is satisfied for γ low enough if

$\frac{1/2 - \hat{p}_B}{(1 - \hat{p}_B)(1 + 1/\hat{p}_B)} < \frac{1}{2 + 1/\hat{p}_B} \Leftrightarrow \hat{p}_B > -1/(2\hat{p}_B)$ which is true. (b) and (d) can be satisfied

simultaneously if $\frac{1 - \gamma/\hat{p}_A}{\gamma(1 + 1/\hat{p}_A)} < \frac{2 - \gamma/\hat{p}_A}{\gamma(2 + 1/\hat{p}_A)}$ which is satisfied for γ low enough

if $\frac{1}{1 + 1/\hat{p}_B} < \frac{1}{1 + 1/2\hat{p}_B}$ which holds. (b) and (c') can be satisfied simultaneously if:

$\frac{(1 - \hat{p}_B)(1 - \gamma/\hat{p}_A)}{(1 - \hat{p}_B)\gamma(1 + 1/\hat{p}_A)} < \frac{1 - 2\hat{p}_B}{\gamma(1 - \hat{p}_B)}$

which is satisfied for γ low enough if: $\frac{1}{1 + 1/\hat{p}_B} < \frac{1 - 2\hat{p}_B}{(1 - \hat{p}_B)} \Leftrightarrow -\hat{p}_B^2 - 2\hat{p}_B + 1 > 0$

$\Leftrightarrow \hat{p}_B < -1 + \sqrt{2}$.

So we have that (a),(b),(c'),(d),(e) can be satisfied if $\hat{p}_B < -1 + \sqrt{2}$.

So we have that for all \hat{p}_B , for all \hat{p}_A close enough to \hat{p}_B , there exists $N_{A\text{low}} < N_{A\text{high}}$ such that if N_B is large enough, then $N_{A\text{low}} < N_A < N_{A\text{high}}$ implies that both players choose B at $t=1$ in any SPE of the game but that this is not socially optimal. Finally, note that if $N_A < N_{A\text{high}}$ then it is not socially optimal for both players to choose B .

Proof of proposition 4:

At period t , player i receives $(1-\gamma) \cdot \sum_{j \neq i} C_{jt}$ from period $(t-1)$, and $\gamma \cdot C_{it}$ directly from period t . Then his or her expected discount payoff in period t is:

$$E[(1-\gamma) \cdot \sum_{j \neq i} C_{jt} + \sum_{\tau=t}^{+\infty} \delta^{\tau-t} [\gamma \cdot C_{i\tau} + (1-\gamma) \delta \cdot \sum_{j \neq i} C_{j\tau+1}]] =$$

$$E[\gamma [\sum_{\tau=t}^{+\infty} \delta^{\tau-t} \cdot C_{i\tau}] + (1-\gamma) (\sum_{\tau=t}^{+\infty} \delta^{\tau-t} (\sum_{j \neq i} C_{j\tau}))]] = E[\sum_{\tau=t}^{+\infty} \delta^{\tau-t} \cdot [\gamma [\gamma_{i\tau} + (1-\gamma) \cdot \sum_{j \neq i} C_{j\tau}]]]$$

which is proportional to the moderator's objective function only if $\gamma=1/2$, in which case the participant's revenue from period t is proportional to the group's contribution in period t .

Proof of proposition 5:

Under $S'(1/2)$, Participant i 's payoff in period τ is proportional to $\pi(i, \tau) = C_{i\tau} +$

$$\sum_{j \neq i} \sum_{t=1}^{\tau-1} \lambda(t, \tau) \cdot [C_{j\tau} - \tilde{C}_{j\tau}(i, t)] = C_{i\tau} + \sum_{j \neq i} (\sum_{t=1}^{\tau-1} \lambda(t, \tau)) \cdot C_{j\tau} - \sum_{j \neq i} \sum_{t=1}^{\tau-1} \lambda(t, \tau) \cdot \tilde{C}_{j\tau}(i, t) = C_{i\tau} + \sum_{j \neq i} C_{j\tau}$$

$- \sum_{j \neq i} \sum_{t=1}^{\tau-1} \lambda(t, \tau) \cdot \tilde{C}_{j\tau}(i, t)$ because $\sum_{t=1}^{\tau-1} \lambda(t, \tau) = 1$. The last term being a constant, the objectives are

aligned. Also if a proportional sharing rule $\beta \cdot \sum_j C_{jt}$ is such that participants search in each pe-

riod, then the participants also search in each period under $S'(1/2)$ with the same β (what matters is the difference between the payoff obtained from searching and the payoff obtained from not searching, which is unaffected), and the total payoff distributed is lower.

Chapter 2: Fast Polyhedral Adaptive Conjoint Estimation

Abstract

We propose and test new adaptive question design and estimation algorithms for partial-profile conjoint analysis. Polyhedral question design focuses questions to reduce a feasible set of parameters as rapidly as possible. Analytic center estimation uses a centrality criterion based on consistency with respondents' answers. Both algorithms run with no noticeable delay between questions.

We evaluate the proposed methods relative to established benchmarks for question design (random selection, D-efficient designs, Adaptive Conjoint Analysis) and estimation (Hierarchical Bayes). Monte Carlo simulations vary respondent heterogeneity and response errors. For low numbers of questions, polyhedral question design does best (or is tied for best) for all tested domains. For high numbers of questions, efficient Fixed designs do better in some domains. Analytic center estimation shows promise for high heterogeneity and for low response errors; Hierarchical Bayes for low heterogeneity and high response errors. Other simulations evaluate hybrid methods, which include self-explicated data.

A field test (330 respondents) compared methods on both internal validity (holdout tasks) and external validity (actual choice of a laptop bag worth approximately \$100). The field test is consistent with the simulation results and offers strong support for polyhedral question design. In addition, marketplace sales were consistent with conjoint-analysis predictions.

1. Polyhedral Methods for Conjoint Analysis

We propose and test (1) a new adaptive question design method that attempts to reduce respondent burden while simultaneously improving accuracy and (2) a new estimation procedure based on centrality concepts. For each respondent the question design method dynamically adapts the design of the next question using that respondent's answers to previous questions. Because the methods make full use of high-speed computations and adaptive, customized local web pages, they are ideally suited for web-based panels. The adaptive method interprets question design as a mathematical program and estimates the solution to the program using recent developments based on the interior points of polyhedra. The estimation method also relies on interior point techniques and is designed to provide robust estimates from relatively few questions. The question design and estimation methods are modular and can be evaluated separately and/or combined with a range of existing methods.

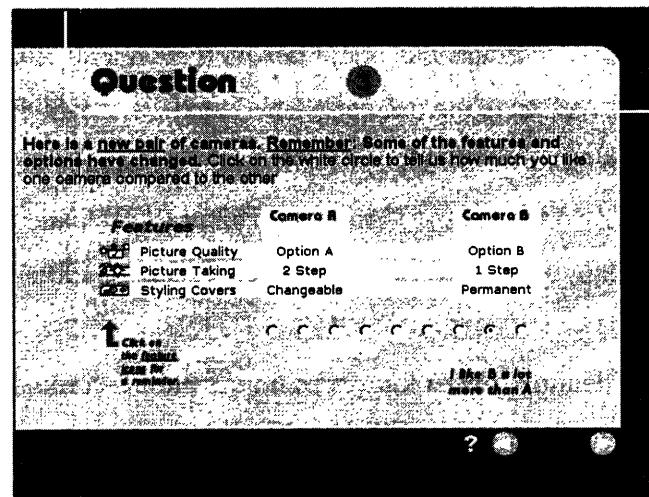
Adapting question design within a respondent, using that respondent's answers to previous questions, is a difficult dynamic optimization problem. Adaptation *within* respondents should be distinguished from techniques that adapt *across* respondents. Sawtooth Software's Adaptive Conjoint Analysis (ACA) is the only published method of which we are aware that attempts to solve this problem (Johnson 1987, 1991). In contrast, aggregate customization methods, such as the Huber and Zwerina (1996), Arora and Huber (2001), and Sandor and Wedel (2001, 2002) algorithms, adapt designs across respondents based on either pretests or Bayesian priors.

ACA uses a data-collection format known as metric paired-comparison questions, and relies on balancing utility between the pairs subject to orthogonality and feature balance. We provide an example of a metric-paired comparison question in Figure 1. To date, aggregate customization methods have focused on a stated-choice data-collection format known as choice-based conjoint (CBC; e.g. Louviere, Hensher, and Swait 2000). Polyhedral methods can be used to design either metric-paired-comparison questions or choice-based questions. In this paper we focus on metric-paired-comparison questions because this is one of the most widely used and applied data-collection formats for conjoint analysis (Green, Krieger and Wind 2001, p. S66; Ter Hofstede, Kim, and Wedel 2002 p. 259). In addition, metric paired-comparison questions are common in computer-aided interviewing, have proven reliable in previous studies (Reibstein,

Bateson, and Boulding 1988; Urban and Katz 1983), provide interval-scaled data with strong transitivity properties (Hauser and Shugan 1980), provide valid and reliable parameter estimates (Leigh, MacKay, and Summers 1984), and enjoy wide use in practice and in the literature (Wit-tink and Cattin 1989). We are extending polyhedral methods to CBC formats (Toubia, Hauser, and Simester, 2003).

Our goal is an initial evaluation of polyhedral methods relative to existing methods under a variety of empirically relevant conditions. We do not expect that any one method will always out-perform the benchmarks, nor do we intend that our findings be interpreted as criticism of any of the benchmarks. Our findings indicate that polyhedral methods have the potential to enhance the effectiveness of existing conjoint methods by providing new capabilities that complement existing methods.

Figure 1
Metric Paired-Comparison Format for I-Zone Camera Redesign



Because the methods are new and adopt a different estimation philosophy, we use Monte Carlo experiments to explore their properties. The Monte Carlo experiments explore the conditions under which polyhedral methods are likely to do better or worse than extant methods. We demonstrate practical domains where polyhedral methods show promise relative to a representative set of widely applied and studied methods. The findings also highlight opportunities for future research by illustrating domains where improvements are necessary and/or where extant methods are likely to remain superior.

We also undertake a large-scale empirical test involving a real product – a laptop computer bag worth approximately \$100. Respondents first completed a series of web-based conjoint questions chosen by one of three question design methods (the methods were assigned randomly). After a filler task, respondents in the study were given \$100 to spend on a choice set of five bags. Respondents received their chosen bag together with the difference in cash between the price of their chosen bag and the \$100. We compare question design and estimation methods on both internal and external validity. Internal validity is evaluated by comparing how well the methods predict several holdout conjoint questions. External validity is evaluated by comparing how well the different conjoint methods predict which bag respondents later chose to purchase using their \$100.

The paper is structured as follows. We begin by describing polyhedral question design and analytic center estimation for metric paired-comparison tasks. Detailed mathematics are provided in Appendix 1 and open-source code is available from <http://mitsloan.mit.edu/vc>. We next describe the design and results of the Monte Carlo experiments. Finally, we describe the field test and the comparative results. We close with a description of the launch of the laptop bag, a summary of the findings, and suggestions for future research.

2. Polyhedral Question Design and Estimation

We begin with a conceptual description that highlights the geometry of the conjoint-analysis parameter space. We illustrate the concepts with a 3-parameter problem because 3-dimensional spaces are easy to visualize and explain. The methods generalize easily to realistic problems that contain ten, twenty, or even one hundred product features. Indeed, relative to existing methods, the polyhedral methods are proposed for larger numbers of product features. By a parameter, we refer to a partworth that needs to be estimated. For example, twenty features with two levels each require twenty parameters because we can scale to zero the partworth of the least preferred feature. Similarly, ten three-level features also require twenty parameters. Interactions among features require still more parameters.

Suppose that we have three features of an instant camera – picture quality, picture taking (2-step vs. 1-step), and styling covers (changeable vs. permanent). If we scale the least desirable level of each feature to zero we have three non-negative parameters to estimate, u_1 , u_2 , and u_3 , reflecting the additional utility (partworth) associated with the most desirable level of

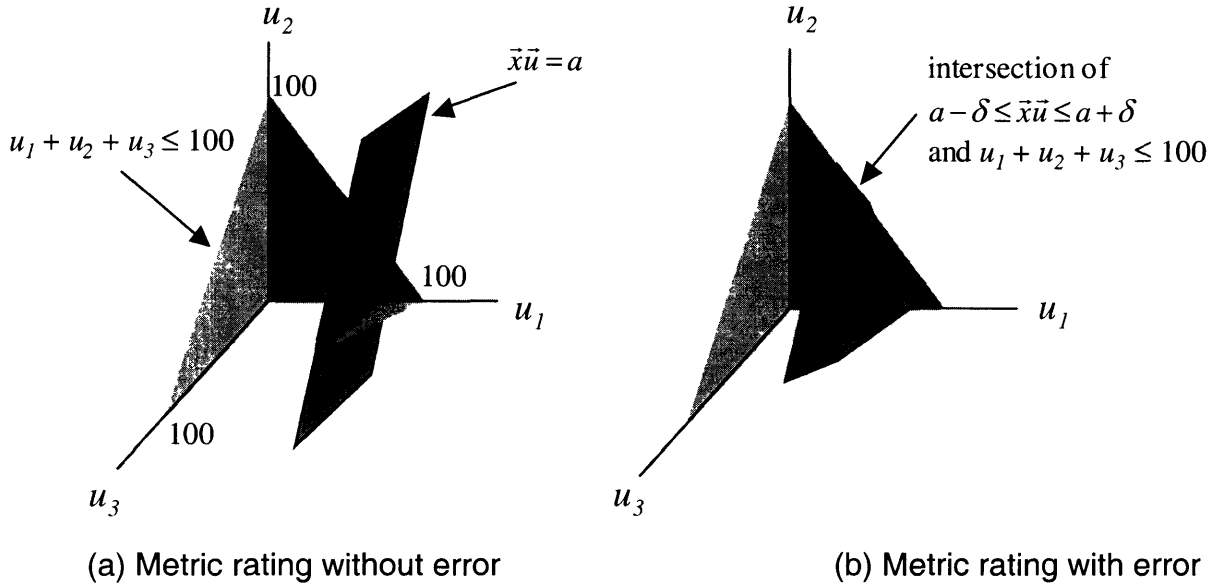
each feature.¹⁹ The measurement scale on which the questions are asked imposes natural boundary conditions. For example, the sum of the partworths of Camera A minus the sum of the partworths of Camera B can be at most equal to the maximum scale difference. In practice, the partworths only have relative meaning and so scaling allows us to impose a wide range of boundary conditions, without loss of generality. Therefore, in order to better visualize the algorithm, we impose a constraint that the sum of the parameters does not exceed some large number (e.g., 100). Under this constraint, prior to any data collection, the feasible region for the parameters is the 3-dimensional bounded polyhedron in Figure 2a.

Suppose that we ask the respondent to evaluate a pair of profiles that vary on one or more features and the respondent says (1) that he or she prefers profile C_1 to profile C_2 and (2) provides a rating, a , to indicate the strength of his or her preference. Assuming for the moment that the respondent answers without error, this introduces an equality constraint that the utility associated with profile C_1 exceeds the utility of C_2 by an amount equal to the rating. If we define $\vec{u} = (u_1, u_2, u_3)^T$ as the 3×1 vector of parameters, \vec{z}_ℓ as the 1×3 vector of product features for the left profile, and \vec{z}_r as the 1×3 vector of product features for the right profile, then, for additive utility, this equality constraint can be written as $\vec{z}_\ell \vec{u} - \vec{z}_r \vec{u} = a$. We can use geometry to characterize what we have learned from this response.

Specifically, we define $\vec{x} = \vec{z}_\ell - \vec{z}_r$ such that \vec{x} is a 1×3 vector describing the difference between the two profiles in the question. Then, $\vec{x} \vec{u} = a$ defines a hyperplane through the polyhedron in Figure 2a. The only feasible values of \vec{u} are those that are in the intersection of this hyperplane and the polyhedron. The new feasible set is also a polyhedron, but it is reduced by one dimension (2-dimensions rather than 3-dimensions). Because smaller polyhedra mean fewer parameter values are feasible, questions that reduce the size of the initial polyhedron as fast as possible lead to more precise estimates of the parameters.

¹⁹ In this example, we assume preferential independence which implies an additive utility function. We can handle interactions by relabeling features. For example, a 2×2 interaction between two features is equivalent to one four-level feature. We hold to this convention throughout the paper.

Figure 2
Respondent's Answers Affect the Feasible Region



However, in any real problem we expect a respondent's answer to contain error. We can model this error as a probability density function over the parameter space (as in standard statistical inference). Alternatively, we can incorporate imprecision in a response by treating the equality constraint $\bar{x}\bar{u} = a$ as a set of two inequality constraints: $a - \delta \leq \bar{x}\bar{u} \leq a + \delta$. In this case, the hyperplane defined by the question-answer pair has "width." The intersection of the initial polyhedron and the "fat" hyperplane is now a three-dimensional polyhedron as illustrated in Figure 2b.

When we ask more questions we constrain the parameter space further. Each question, if asked carefully, will result in a hyperplane that intersects a polyhedron resulting in a smaller polyhedron – a "thin" region in Figure 2a or a "fat" region in Figure 2b. Each new question-answer pair slices the polyhedron in Figure 2a or 2b yielding more precise estimates of the parameter vector \bar{u} .

We incorporate prior information about the parameters by imposing constraints on the parameter space. For example, if u_m and u_h are the medium and high levels, respectively, of a feature, then we impose the constraint $u_m \leq u_h$ on the polyhedron. Previous research suggests that these types of constraints enhance estimation (Johnson 1999; Srinivasan and Shocker 1973). We

now examine question design for metric paired-comparison data by dealing first with the case in which subjects respond without error (Figure 2a). We then describe how to modify the algorithm to handle error (e.g., Figure 2b).

Selecting Questions to Shrink the Feasible Set Rapidly

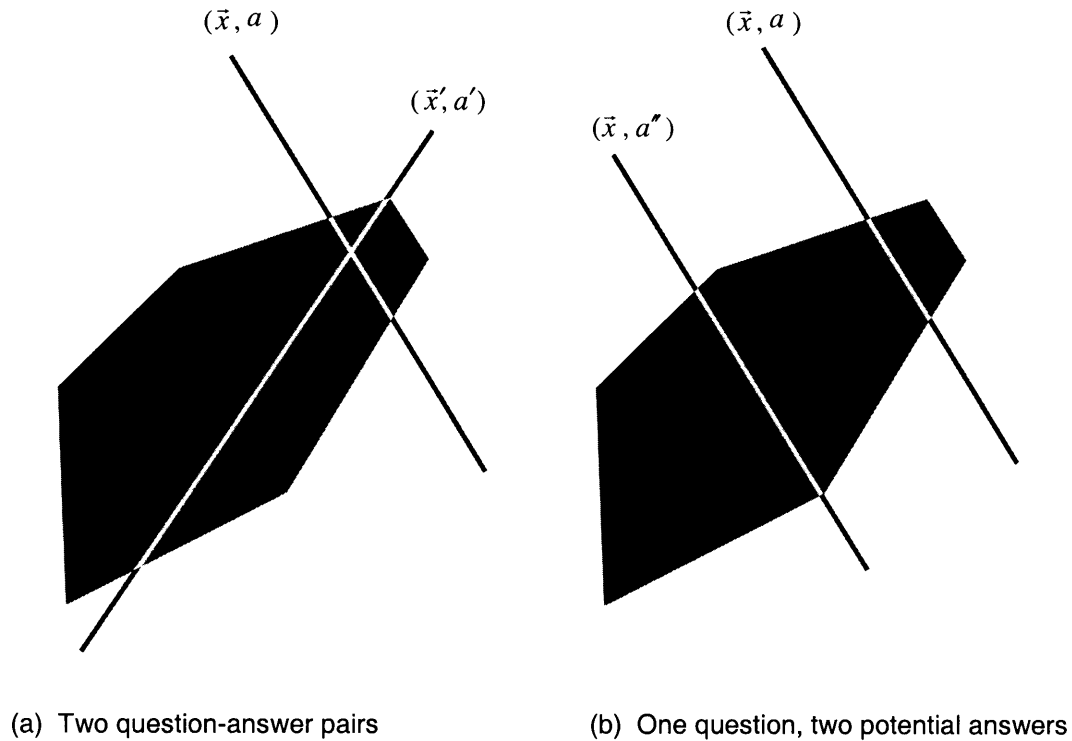
The question design task describes the design of the profiles that respondents are asked to compare. Questions are more informative if the answers allow us to estimate partworths more quickly. For this reason, we select the respondent's next question in a manner that is likely to reduce the size of the feasible set (for that respondent) as fast as possible.

Consider for a moment a 20-dimensional problem (without errors in the answers). As in Figure 2a, a question-based constraint reduces the dimensionality by one. That is, the first question reduces a 20-dimensional set to a 19-dimensional set; the next question reduces this set to an 18-dimensional set and so on. After the twelfth question, for example, we reach an 8-dimensional set: 8 dimensions = 20 parameters – 12 questions. Without further restriction, the feasible parameters are generally not unique – any point in the 8-dimensional set (polyhedron) is still feasible. However, the 8-dimensional set might be quite small and we might have a very good idea of the partworths. For example, the first twelve questions might be enough to tell us that some features, say picture quality, styling covers, and battery life, have large partworths while other features, say folding capability, light selection, and film ejection method, have very small partworths. If this holds across respondents then, during an early phase of a product development process, the product development team might feel they have enough information to focus on the key features.

Although the polyhedral algorithm is designed for high-dimensional spaces, it is hard to visualize 20-dimensional polyhedra. Instead, we illustrate the polyhedral question design method in a situation where the remaining feasible set is easy to visualize. Specifically, by generalizing our notation slightly to q questions and p parameters, we define \bar{a} as the $q \times 1$ vector of answers and X as the $q \times p$ matrix with rows equal to \bar{x} for each question (recall that \bar{x} is a $1 \times p$ vector). Then the respondent's answers to the first q questions define a $(p-q)$ -dimensional hyperplane given by the equation $X\bar{u} = \bar{a}$. This hyperplane intersects the initial p -dimensional polyhedron to give us a $(p-q)$ -dimensional polyhedron. In the example of $p=20$ parameters and $q=18$

questions, the result is a 2-dimensional polyhedron that is easy to visualize. One such 2-dimensional polyhedron is illustrated in Figure 3.

Figure 3
Choice of Question (2-dimensional slice)



Our task is to select questions that reduce the 2-dimensional polyhedron as fast as possible. Mathematically, we select a new question vector, \bar{x} , and the respondent answers this question with a new rating, a . We add the new question vector as the last row of the question matrix and we add the new answer as the last row of the answer vector. While everything is really happening in p -dimensional space, the net result is that the new hyperplane will intersect the 2-dimensional polyhedron in a line segment (i.e., a 1-dimensional polyhedron). The slope of the line will be determined by \bar{x} and the intercept by a . We illustrate two potential question-answer pairs in Figure 3a. The slope of the line is determined by the question, the specific line by the answer, and the remaining feasible set by the line segment within the polyhedron. In Figure 3a one of the question-answer pairs (\bar{x}, a) reduces the feasible set more rapidly than the other question-answer pair (\bar{x}', a') . Figure 3b repeats a question-answer pair (\bar{x}, a) and illustrates an alternative answer to the same question (\bar{x}, a'') .

If the polyhedron is elongated as in Figure 3, then, in most cases, questions that imply line segments perpendicular to the longest “axis” of the polyhedron are questions that result in the smallest remaining feasible sets. Also, because the longest “axis” is in some sense a bigger target, it is more likely that the respondent’s answer will select a hyperplane that intersects the polyhedron. From analytic geometry we know that hyperplanes (line segments in Figure 3) are perpendicular to their defining vectors (\vec{x}). Thus, we can reduce the feasible set as fast as possible (and make it more likely that answers are feasible) if we choose question vectors that are parallel to the longest “axis”. For example, both line segments based on \vec{x} in Figure 3b are shorter than the line segment based on \vec{x}' in Figure 3a.

If we can develop an algorithm that works in any p -dimensional space, then we can generalize this intuition to any question, q , such that $q \leq p$. After receiving answers to the first q questions, we could find the longest vector of the $(p-q)$ -dimensional polyhedron of feasible parameter values. We could then ask the question based on a vector that is parallel to this “axis.” The respondent’s answer creates a hyperplane that intersects the polyhedron to produce a new polyhedron. We address later the cases where respondents’ answers contain error and where $q > p$.

Centrality Estimation

Polyhedral geometry also gives us a means to estimate the parameter vector, \vec{u} , when $q \leq p$. Recall that, after question q , any point in the remaining polyhedron is consistent with the answers the respondent has provided. If we impose a diffuse prior that any feasible point is equally likely, then we would like to select the point that minimizes the expected error. This point is the center of the feasible polyhedron, or more precisely, the polyhedron’s center of gravity. The smaller the feasible set, either due to better question design or more questions (higher q), the more precise the estimate. If there were no respondent errors, then the estimate would converge to its true value when $q=p$ (the feasible set becomes a single point, with zero dimensionality). For $q > p$ the same point would remain feasible. As we discuss below, this changes when responses contain error.

This technique of estimating partworths from the center of a feasible polyhedron is related to that proposed by Srinivasan and Shocker (1973, p. 350) who suggest using a linear program to find the “innermost” point that maximizes the minimum distance from the hyperplanes that bound the feasible set. Philosophically, the proposed polyhedral method makes maximum

use of the information in the constraints and then takes a central estimate based on what is still feasible. Carefully chosen questions shrink the feasible set rapidly. We then use a centrality estimate that has proven to be a surprisingly good approximation in a variety of engineering problems. More generally, the centrality estimate is similar in some respects to the proven robustness of linear models, and in some cases, to the robustness of equally-weighted models (Dawes and Corrigan 1974; Einhorn 1971, Huber 1975; Moore and Semenik 1988; Srinivasan and Park 1997).

Interior-Point Algorithms and the Analytic Center of a Polyhedron

To select questions and obtain intermediate estimates the proposed heuristics require that we solve two non-trivial mathematical programs. First, we must find the longest “axis” of a polyhedron (to select the next question) and second, we must find the polyhedron’s center of gravity (to provide a centrality estimate). If we were to define the longest “axis” of a polyhedron as the longest line segment in the polyhedron, then one method to find the longest “axis” would be to enumerate the vertices of the polyhedron and compute the distances between the vertices. However, solving this problem requires checking every extreme point, which is computationally intractable (Gritzmann and Klee 1993). In practice, solving the problem would impose noticeable delays between questions. Also, the longest line segment in a polyhedron may not capture the concept of a longest “axis.” Finding the center of gravity of the polyhedron is even more difficult and computationally demanding.

Fortunately, recent work in the mathematical programming literature has led to extremely fast algorithms based on projections within the interior of polyhedrons (much of this work started with Karmarkar 1984). Interior-point algorithms are now used routinely to solve large problems and have spawned many theoretical and applied generalizations. One such generalization uses bounding ellipsoids. In 1985, Sonnevend demonstrated that the shape of a bounded polyhedron can be approximated by proportional ellipsoids, centered at the “analytic center” of the polyhedron. The analytic center is the point in the polyhedron that maximizes the geometric mean of the distances to the boundaries of the polyhedron. It is a central point that approximates the center of gravity of the polyhedron, and finds practical use in engineering and optimization. Furthermore, the axes of the ellipsoids are well-defined and intuitively capture the concept of an

“axis” of a polyhedron. For more details see Freund (1993), Nesterov and Nemirovskii (1994), Sonnevend (1985a, 1985b), and Vaidja (1989).

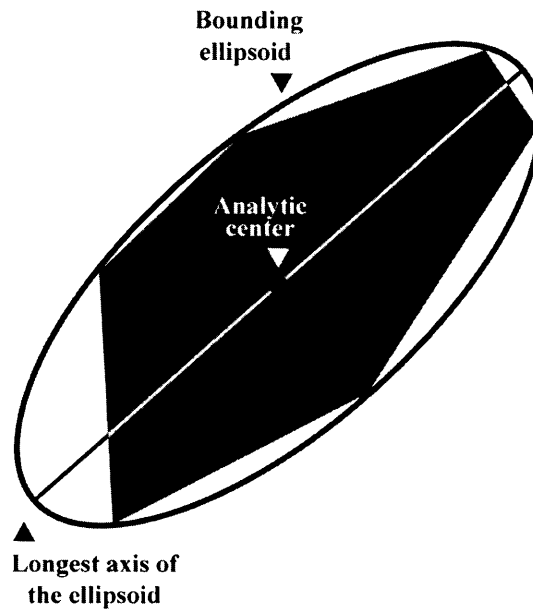
Polyhedral Question Design and Analytic Center Estimation

We illustrate the proposed process in Figure 4, using the same two-dimensional polyhedron depicted in Figure 3. The algorithm proceeds in four steps. We first find a point in the interior of the polyhedron. This is a simple linear programming (LP) problem and runs quickly. Then, following Freund (1993) we use Newton’s method to make the point more central. This is a well-formed problem and converges quickly to yield the analytic center as illustrated by the black dot in Figure 4. We next find a bounding ellipsoid based on a formula that depends on the analytic center and the question-matrix, X . We then find the longest axis of the ellipsoid (diagonal line in Figure 4) with a quadratic program that has a closed-form solution. The next question, \bar{x} , is based on the vector most nearly parallel to this axis. A formal (mathematical) description of each step is provided in Appendix 1.

Analytically, this algorithm works well in higher dimensional spaces. For example, Figure 5 illustrates the algorithm when $(p - q) = 3$, where we reduce a 3-dimensional feasible set to a 2-dimensional feasible set. Figure 5a illustrates a polyhedron based on the first q questions. Figure 5b illustrates a bounding 3-dimensional ellipsoid, the longest axis of that ellipsoid, and the analytic center. The longest axis defines the question that is asked next which, in turn, defines the slope of the hyperplane that intersects the polyhedron. One such hyperplane is shown in Figure 5c. The respondent’s answer locates the specific hyperplane. The intersection of the selected hyperplane and the 3-dimensional polyhedron is a new 2-dimensional polyhedron, such as that in Figure 4. This process applies (in higher dimensions) from the first question to the p^{th} question. For example, the first question implies a hyperplane that cuts the first p -dimensional polyhedron such that the intersection yields a $(p - 1)$ -dimensional polyhedron.

The polyhedral algorithm runs extremely fast. We have implemented the algorithm for the web-based empirical test described later in this paper. Based on this example, with ten two-level features, respondents noticed no delay in question design nor any difference in speed versus a fixed design. For a demonstration see the website referenced earlier.

Figure 4
Bounding Ellipsoid and the Analytic Center (2-dimensions)

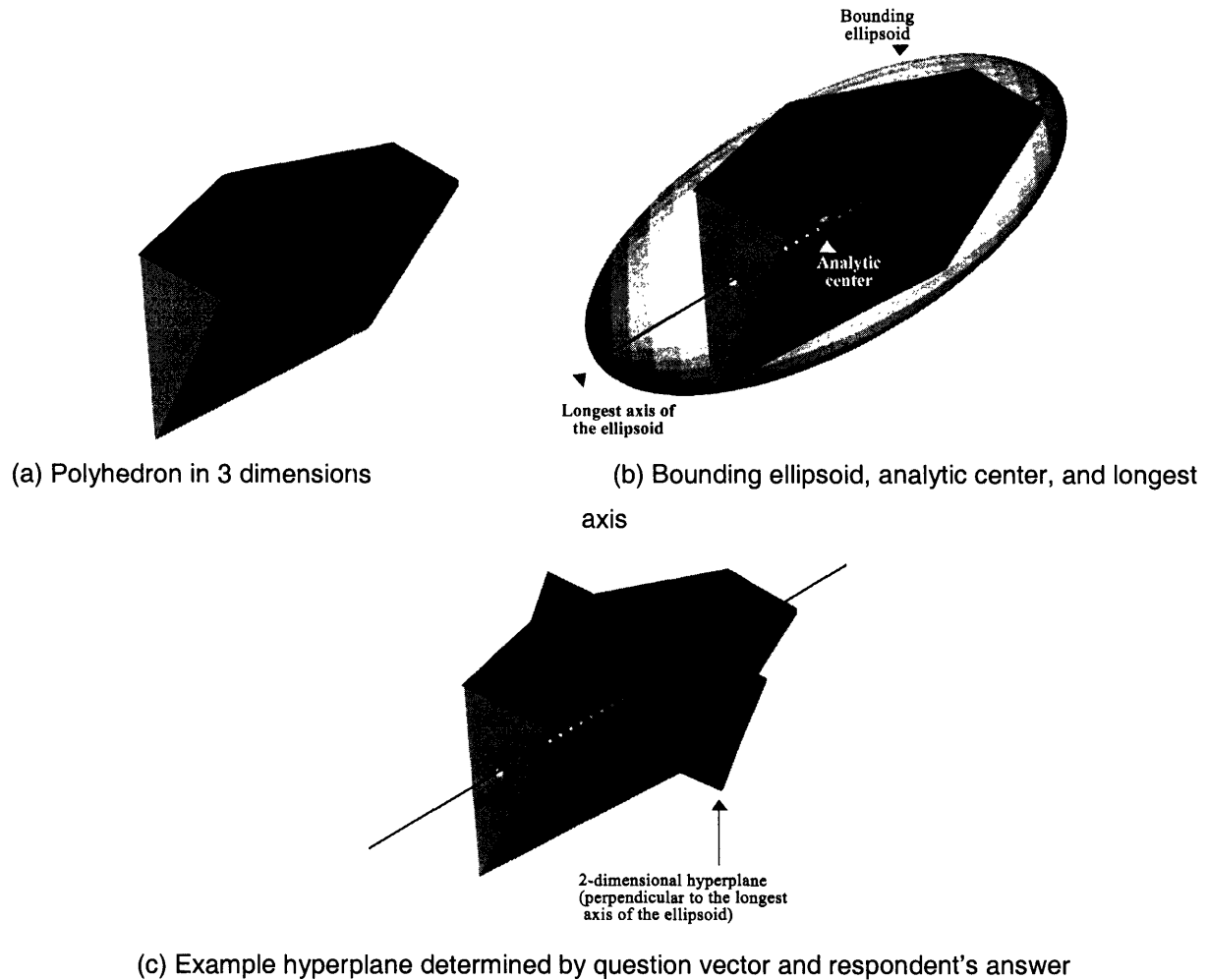


Inconsistent Responses and Error-Modeling

Figures 2, 3, 4, and 5 illustrate the geometry when respondents answer without error. However, real respondents are unlikely to be perfectly consistent. It is more likely that, for some $q < p$, the respondent's answers will be inconsistent and the polyhedron will become empty. That is, we will no longer be able to find any parameters, \bar{u} , that satisfy the equations that define the polyhedron, $X\bar{u} = \bar{a}$. Thus, for real applications, we extend the polyhedral algorithm to address response errors. Specifically, we adjust the polyhedron in a minimal way to ensure that some parameter values are still feasible. We do this by modeling errors, $\bar{\delta}$, in the respondent's answers such that $\bar{a} - \bar{\delta} \leq X\bar{u} \leq \bar{a} + \bar{\delta}$ (recall Figure 2b). We then choose the minimum errors such that these constraints are satisfied. This same modification covers estimation for the case of $q > p$. Appendix 1 provides the mathematical program (OPT4) that we use to estimate \bar{u} and $\bar{\delta}$. The algorithm is easily modified to incorporate alternative error formulations, such as least-

squares or minimum sum of absolute deviations, rather than this “minimax” criterion.²⁰ Exploratory simulations suggest that the algorithm is robust to the choice of error criterion.

Figure 5
Question design with a 3-Dimensional Polyhedron



To implement this policy for analytic center estimation, we use a two-stage algorithm. In the first stage we treat the responses as if they occurred without error – the feasible polyhedron shrinks rapidly and the analytic center is a working estimate of the true parameters. However, as soon as the feasible set becomes empty, we adjust the constraints by adding or subtracting “er-

²⁰ Technically, the minimax criterion is called the “ ∞ -norm.” To handle least-squares errors we use the “2-norm” and to handle average absolute errors we use the “1-norm.” Either is a simple modification to OPT4 in Appendix 1.

rors,” where we choose the minimum errors, $\|\bar{\delta}\|$, for which the feasible set is non-empty. The analytic center of the new polyhedron becomes the working estimate and $\bar{\delta}$ becomes an index of response error.²¹ As with all of our heuristics, the accuracy of our error-modeling method is tested with simulation. While estimates based on this heuristic seems to converge for the domains that we test (reduce mean errors, no measured bias), we recognize the need for further exploration of this heuristic, together with the development of a formal error theory.

Addressing Other Practical Implementation Issues

Implementation raises several additional issues. Alternative solutions to these issues may yield more or less accurate parameter estimates, and so the performance of the polyhedral methods in the validation tasks are lower bounds on the performance of this class of methods.

Product profiles with discrete features. In most conjoint analysis problems the features are specified at discrete levels, as in Figure 1. This constrains the elements of the \bar{x} vector to be 1, -1, 0, or 0, depending on whether the left profile, the right profile, neither profile, or both profiles have the “high” feature, respectively. In this case we choose the vector that is most nearly parallel to the longest axis of the ellipsoid. Because we can always recode multi-level features or interacting features as binary features, the geometric insights still hold even if we otherwise simplify the algorithm.

Restrictions on question design. For a p -dimensional problem we may wish to vary fewer than p features in any paired-comparison question. For example, Sawtooth Software (1996, p. 7) suggests that: “Most respondents can handle three attributes after they’ve become familiar with the task. Experience tells us that there does not seem to be much benefit from using more than three attributes.” We incorporate this constraint by restricting the set of questions over which we search when finding a question-vector that is parallel to the longest axis of the ellipse.

First question. Unless we have prior information before any question is asked, the initial polyhedron of feasible utilities is defined by the boundary constraints. If the boundary constraints are symmetric, the polyhedron is also symmetric and the polyhedral method offers

²¹ By construction, $\bar{\delta}$ grows (weakly) with the number of questions, q , thus a better measure of fit might be mean error, the error divided by the number of questions – an analogy to mean-squared error in regression.

little guidance for the choice of the first question. In these situations we choose the first question for each respondent so that it helps improve estimates of the population means by balancing how often each feature level appears in the set of questions answered by all respondents. In particular, for the first question presented to each respondent we choose feature levels that appeared infrequently in the questions answered by previous respondents.

Question design when the parameter set becomes infeasible. Analytic center estimation is well-defined when the parameter set becomes infeasible, but question design is not. Thus, in the simulations we use a random question design heuristic when the parameter set is infeasible.²² This provides a lower bound on what might be achieved.

Programming. The optimization algorithms used for the simulations are written in Matlab and are available at the website cited earlier. We also provide the simulation code and demonstrations of web-based applications. All code is open-source.

3. Monte Carlo Simulations

Polyhedral methods for conjoint analysis are new and untested. Although interior-point algorithms and the centrality criterion have been successful in many engineering problems, we are unaware of any prior application to marketing problems. Thus, we turn first to Monte Carlo experiments to identify circumstances in which polyhedral methods may contribute to the effectiveness of current methods. Monte Carlo simulations offer at least three advantages for the initial test of a new method. First, they facilitate comparison of different techniques in a range of domains such as varying levels of respondent heterogeneity and response accuracy. We can also evaluate combinations of the techniques, for example, mixing polyhedral question design with extant estimation methods. Second, simulations resolve the issue of identifying the correct answer. In studies involving actual customers, the true partial utilities are unobserved. In simulations the true partial utilities are constructed so that we can compare how well alternative methods identify the true utilities from noisy responses. Finally, other researchers can readily replicate the findings. However, simulations do not guarantee that real respondents behave as simu-

²² Earlier implementations, including the field test, used ACA question design when the parameter set became infeasible. Further analysis revealed that it is better to switch to random question design than ACA question design when the parameter set becomes infeasible. Fortunately, this makes the performance of polyhedral question design in the field test conservative.

lated nor do they reveal which domain is likely to best summarize field experience. Thus, following the simulations, we examine a field test that matches one of the simulated domains.

Many papers have used the relative strengths of Monte Carlo experiments to study conjoint techniques, providing insights on interactions, robustness, continuity, feature correlation, segmentation, new estimation methods, new data-collection methods, post-analysis with Hierarchical Bayes methods, and comparisons of ACA, CBC, and other conjoint methods. Although we focus on specific benchmarks, there are many comparisons in the literature of these benchmarks to other methods (see reviews and citations in Green 1984; Green, Krieger, and Wind 2001, 2002; Green and Srinivasan 1978, 1990; Hauser and Rao 2003; Moore 2003.)

We test polyhedral question design versus three question design benchmarks and analytic center estimation versus Hierarchical Bayes estimation. The initial simulations vary respondent heterogeneity, accuracy of respondent answers, and the number of questions. In a second set of simulations we also consider the role of self-explicated responses and vary the accuracy of self-explicated responses.

Respondent Heterogeneity, Response Errors, and Number of Questions

We focus on a design problem involving ten features, where a product development team is interested in learning the incremental utility contributed by each feature. We follow convention and scale to zero the partworth of the low level of a feature and, without loss of generality, bound it by 100. This results in a total of ten parameters to estimate ($p = 10$). We anticipate that the polyhedral methods are particularly well-suited to solving problems in which there are a large number of parameters relative to the number of responses from each individual ($q < p$). Thus, we vary the number of questions from slightly less than the number of parameters ($q = 8$) to comfortably more than the number of parameters ($q = 16$).

We simulate each respondent's partworths by drawing independently and randomly from a normal distribution with mean 50 and variance σ_u^2 , truncated to the range. We explored the sensitivity of the findings to this specification by testing different methods of drawing partworths, including beta distributions that tend to yield more similar partworths (inverted-U shape distributions), more diverse partworths (U-shaped distributions), or moderately diverse partworths (uniform distributions). Sensitivity analyses for key findings did not suggest much variation. Nonetheless, this is an important area for more systematic future research. By manipulat-

ing the standard deviation of the normal distribution we explore a relatively homogeneous population ($\sigma_u = 10$) and a relatively heterogeneous population ($\sigma_u = 30$). These values were chosen because they are comparable to those used elsewhere in the literature, because they result in moderate-to-low truncation, and because their range illustrates how the accuracy of the methods varies with heterogeneity.

To simulate the response to each metric paired-comparison (PC) question, we calculate the true utility difference between each pair of product profiles by multiplying the design vector by the vector of true partworths: $\bar{x}\bar{u}$. We assume that the respondents' answers to the questions equal the true utility difference plus a zero-mean normal response error with variance σ_{pc}^2 . The assumption of normally distributed error is common in the literature and appears to be a reasonable assumption about PC response errors (Wittink and Cattin 1981 report no systematic effects due to the type of error distribution assumed). We select response errors, comparable to those used in the literature. Specifically, to illustrate the range of response errors we use both a low response error ($\sigma_{pc} = 20$) and a high response error ($\sigma_{pc} = 40$).²³ For each comparison, we simulate 500 respondents (in five sets of 100).

Question Design Benchmarks

We compare the Polyhedral question design method against three benchmarks: Random question design, a Fixed design, and the question design used by Adaptive Conjoint Analysis (ACA).²⁴ For the Random benchmark, the feature levels are chosen randomly and equally-likely. The Fixed design provides another non-adaptive benchmark. For $q > p$, we select the ($q = 16$) design with an algorithm that seeks the highest obtainable D-efficiency (Kuhfield, Tobias, and Garratt 1994). Efficiency is not defined for $q < p$, thus, for $q = 8$, we follow the procedure established by Lenk, et. al. (1996) and choose questions randomly from an efficient design for $q = 16$.

We choose ACA question design as our third benchmark because it is the industry and academic standard for within-respondent adaptive question design. For example, in 1991 Green,

²³ Response errors in the literature, often reported as a ratio of error variance to true variance (heterogeneity), vary considerably. In our case, the "respondent's" answer, a , is the difference between the sum of the u_f 's. Thus the variance of a is a multiple of σ_u^2 . For our situation, the percent errors vary from 8% to 57%.

²⁴ Beginning in this section we capitalize the question design and estimation methods for easy reference. We retain lower case for generic descriptions.

Krieger and Agarwal (p. 215) stated that “in the short span of five years, Sawtooth Software’s Adaptive Conjoint Analysis has become one of the industry’s most popular software packages for collecting and analyzing conjoint data,” and go on to cite a number of academic papers on ACA. Although accuracy claims vary, ACA appears to predict reasonably well in many situations (Johnson 1991; Orme 1999).

The ACA method includes five sections: an unacceptability task (that is often skipped), a ranking of the features, a series of self-explicated (SE) questions, the metric paired-comparison (PC) questions, and purchase intentions for calibration concepts. The question design procedure, has not changed since it was “originally programmed for the Apple II computer in the late 70s” (Orme and King 2002). It adapts the PC questions based on intermediate estimates (after each question) of the partworths. These intermediate estimates are based on an OLS regression using the SE and PC responses and ensure that the pairs of profiles are nearly equal in estimated utility (utility balance). Additional constraints restrict the overall design to be nearly orthogonal (features and levels are presented independently) and balanced (features and levels appear with near equal frequency).

To avoid handicapping the ACA question design in the initial simulations, we simulate the SE responses without adding error. In particular, Sawtooth Software asks for SE responses using a 4-point scale, in which the respondent states the relative importance of improving the product from one feature level to another (e.g., adding automatic film ejection to an instant camera). We set the SE responses equal to the true partworths, but discretize the answer to match the ACA scale.

Our code was written using Sawtooth Software’s documentation together with e-mail interactions with the company’s representatives. We then confirmed the accuracy of the code by asking Sawtooth Software to re-estimate partworths for a small sample of data.

Estimation Benchmark

The two estimation methods are the Analytic Center (AC) method described earlier and Hierarchical Bayes (HB) estimation. Hierarchical Bayes estimation uses data from the population to inform the distribution of partworths across respondents and, in doing so, estimates the posterior mean of respondent-level partworths with an algorithm based on Gibbs sampling and the Metropolis Hastings Algorithm (Allenby and Rossi 1999; Arora, Allenby and Ginter 1998; John-

son 1999; Lenk, et. al. 1996; Liechty, Ramaswamy and Cohen 2001; Sawtooth Software 2001; Yang, Allenby and Fennell 2002). For ACA question design, Sawtooth Software recommends HB as their most accurate estimation method (Sawtooth Software 2002, p. 11). For this initial comparison, for all question design methods, we use data from the SEs as starting values and we use the SEs to constrain the rank-order of the levels for each feature (Sawtooth Software 2001, p. 13).²⁵

Criterion

To compare the performance of each benchmark we calculate the mean absolute accuracy of the parameter estimates (true vs. estimated values averaged across parameters and respondents). We chose to report mean absolute error (MAE) rather than root mean squared error (RMSE) because the former is less sensitive to outliers and is more robust over a variety of induced error distributions (Hoaglin, Mosteller and Tukey 1983; Tukey 1960). However, as a practical matter, the qualitative implications of our simulations are the same for both error measures. Indeed, except for a scale change, the results are almost identical for both MAE and RMSE. This is not surprising; for normal distributions the two measures differ only by a factor of $(2/\pi)^{1/2}$.

The results are based on the average of five simulations, each with 100 respondents. To reduce unnecessary variance among question design methods, we first draw the partworths and then use the same partworths to evaluate each question design method. The use of multiple draws makes the results less sensitive to spurious effects from a single draw.

4. Results of the Initial Monte Carlo Experiments

We begin with the results obtained from using eight ($q = 8$) paired comparison questions. This is the type of domain for which Polyhedral question design and Analytic Center estimation were developed (more parameters to estimate than there are questions). Moreover, within this domain there are generally a range of partworths that are feasible, and so the polyhedron is not empty. In our simulations the polyhedron contains feasible answers for an average of 7.97 questions when response errors are low. When response errors are high, this average drops to 6.64.

²⁵ Another version of Sawtooth Software's HB algorithm also uses the SEs to constraint the relative partworths across features. We test this version in our next set of simulations. This enables us to isolate the impact of the paired-comparison question design algorithm.

Table 1 reports the MAE in the estimated partworths for a complete crossing of question design methods, estimation methods, response error, and heterogeneity. The best results (lowest error) in each column are indicated by **bold text**. In Table 2 we reorganize the data to indicate the directional impact of either heterogeneity or response errors on the performance of the question design and estimation methods. In particular, we average the performance of each question design method across estimation methods (and vice versa). To indicate the directional effect of heterogeneity we average across response errors (and vice versa).

Table 1
Comparison of Question Design and Estimation Methods for $q = 8$
Mean Absolute Errors

Question design	Estimation	Homogeneous Population		Heterogeneous Population	
		Low Response Error	High Response Error	Low Response Error	High Response Error
Random	AC	16.5	24.1	15.9	21.7
	HB	8.1	10.2	19.8	22.2
Efficient Fixed	AC	13.7	22.9	14.3	21.0
	HB	7.8*	10.3	20.4	22.5
ACA	AC	14.9	24.2	16.1	22.1
	HB	8.3	9.8*	23.9	22.9
Polyhedral	AC	10.7	20.9	12.5*	19.7*
	HB	7.8*	9.9*	20.6	22.2

Smaller numbers indicate better performance.

*For each column, lowest error or not significantly different from lowest ($p < 0.05$). All others are significantly different from lowest.

Table 2
Directional Implications of Response Errors and Heterogeneity for $q = 8$
Mean Absolute Errors

	Homogeneous Population	Heterogeneous Population	Low Response Error	High Response Error
Question design				
Random	14.7	19.9	15.1	19.5
Efficient Fixed	13.6	19.5	14.0	19.1
ACA	14.3	21.2	15.8	19.8
Polyhedral	12.3*	18.4*	12.9*	18.2*
Estimation				
AC	18.5	17.9*	14.3*	22.1
HB	9.0*	21.8	14.6*	16.2*

Smaller numbers indicate better performance.

*For each column within question design or estimation, lowest error or not significantly different from lowest ($p < 0.05$). All others are significantly different from lowest.

Question Design Methods

The findings indicate that when there are only a small number of PC questions, the Polyhedral question design method performs well compared to the other three benchmarks. This conclusion holds across the different levels of response error and heterogeneity. The improvement over the Random question design method is reassuring, but perhaps not surprising. The improvement over the Fixed method is also not surprising when there are a small number of questions, as it is not possible to achieve the balance and orthogonality goals that the fixed method seeks.

The comparison with ACA question design is more interesting. Further investigation reveals that the relatively poor performance of the ACA method can be attributed, in part, to endogeneity bias, resulting from utility balance – the method that ACA uses to adapt questions. To understand this result we first recognize that any adaptive question design method is, potentially, subject to endogeneity bias. Specifically, the q th question depends upon the answers to the first $q-1$ questions. This means that the q th question depends, in part, on any response errors in the first $q-1$ questions. This is a classical problem, which often leads to bias (see for example Judge, et. al. 1985, p. 571). Thus, adaptivity represents a tradeoff: we get better estimates more quickly,

but with the risk of endogeneity bias. In our simulations, the absolute bias with ACA questions and AC estimation is approximately 6.6% of the mean when averaged across domains. This is statistically significant, in part, because of the large sample size in the simulations. Polyhedral question design is also adaptive and it, too, could lead to biases. However, in all four domains, the bias for ACA questions is significantly larger than the bias for Polyhedral questions (1.0% on average for AC). The endogeneity bias in ACA questions appears to be from utility-balanced question design; it is not removed with HB estimation. While further analyses of endogeneity bias are beyond the scope of this paper, they represent an interesting topic for future research. In particular, it might be possible to derive estimation methods that correct for these endogeneity biases.

Estimation Methods

For homogeneous populations, Hierarchical Bayes consistently performed better than Analytic Center estimation, irrespective of the question design method. The performance differences were generally large. Hierarchical Bayes estimation uses population-level data to moderate individual estimates. If the population is homogenous, then, at the individual level, the ratio of noise to true variation is higher and so moderating this variance through population-level data improves accuracy. However, if the population is heterogeneous, then reliance on population data makes it more difficult to identify the true individual-level variation and Analytic Center estimation does better. For a heterogeneous population, the combination of Polyhedral question design and Analytic Center estimation was significantly more accurate than any other combination of question design or estimation method (Table 1).

The findings also suggest that Hierarchical Bayes is relatively more accurate when response errors are high, while Analytic Center estimation is more likely to be favored when response errors are low. The reliance of Hierarchical Bayes on population-level data may also explain the role of response errors. If response errors are large, much of the individual-level variance is due to noise. Population-level data are less sensitive to response errors (due to aggregation) and so reliance on this data helps to improve accuracy. On the other hand, when response errors are low, the polyhedron stays feasible longer and the Analytic Center method appears to do a better job of identifying individual-level variation.

Additional Paired-Comparison Questions

Although polyhedral methods were developed primarily for situations with only a relatively small number of questions, there remain important applications in which a larger number of questions can be asked of each respondent. To examine whether the potential accuracy advantages of polyhedral methods for low q leads to a loss of accuracy at high q , we re-examined the performance of each method after sixteen paired-comparison questions ($q = 16$).

Recall that the Polyhedral method is only used to design questions when the polyhedron contains feasible responses. For low response errors the polyhedron is typically empty after 8 questions, while for high response errors this generally occurs at around 6 or 7 questions. Once the polyhedron is empty we choose questions randomly. Because the Polyhedral question design method is only responsible for around half of the questions we use the label, “Poly/Random.”

The findings are reported in Table 3, which is analogous to the previous Table 2. They reveal the emergence of Fixed question design methods in some domains. Asking a larger number of questions results in more complete coverage of the parameter space. This increases the importance of orthogonality and balance – the criteria used in efficient Fixed question design. With more complete coverage, the ability to customize questions to focus on specific regions of the question space becomes less important, mitigating the advantage offered by adaptive techniques.

However, even after sixteen questions there remain domains in which Polyhedral question design can improve performance. The Poly/Random method appears to be at least as accurate as the Fixed design when the population is homogenous and/or response errors are high. Its advantage for low q does not seem to be particularly harmful for high q , especially for high response errors. Table 3 also suggests that Analytic Center estimation remains a useful estimation procedure when populations are heterogeneous and/or response errors are low. We note that this result is consistent with Andrews, Ansari, and Currim (2002, p. 87) who conclude “individual-level models overfit the data.” They test OLS rather than the Analytic Center method and do not test adaptive methods.

Table 3
Directional Implications of Response Errors and Heterogeneity for $q = 16$
Mean Absolute Errors

	Homogeneous Population	Heterogeneous Population	Low Response Error	High Response Error
Question design				
Random	12.1	14.3	10.4	15.9
Efficient Fixed	10.3	13.1*	8.8*	14.6
ACA	12.5	18.1	13.5	17.2
Poly/Random	9.2*	15.2	10.4	14.0*
Estimation Method				
AC	13.9	12.5*	9.9*	16.4
HB	8.2*	17.9	11.6	14.4*

Smaller numbers indicate better performance.

*For each column within question design or estimation, lowest error or not significantly different from lowest ($p < 0.05$). All others are significantly different from the lowest.

In summary, our Monte Carlo experiments suggest that there are domains in which Polyhedral question design and/or Analytic Center estimation improve the accuracy of conjoint analysis, but there are also domains better served by extant methods. Specifically,

- Polyhedral question design shows promise for low number of questions, such as the fuzzy front-end of product development and/or web-based interviewing.
- For larger numbers of questions, efficient Fixed designs appear to be best, but Poly/Random question design does well, especially when response errors are high and populations are homogenous.
- Analytic Center estimation shows promise for heterogeneous populations and/or low response errors where the advantage of an individual-respondent focus is strongest.
- Hierarchical Bayes estimation is preferred when populations are more homogeneous and response errors are large.

5. The Role of Self-Explicated Questions

Hybrid conjoint models refer to methods that combine both compositional methods, such as self-explicated (SE) questions, and decompositional methods, such as metric paired-comparison (PC) questions, to produce new estimates. Although there are instances in which

methods that use just one of these data sources outperform or provide equivalent accuracy to hybrid methods, there are many situations and product categories in which hybrid methods improve accuracy (e.g., Green 1984).

An important hybrid from the perspective of evaluating polyhedral methods is ACA – the most widely used method for adaptive metric paired-comparison questions. While ACA’s question design algorithm has remained constant since the late 1970s, its estimation procedures have evolved to address the incommensurability of the SE and PC scales. Its default estimation procedure relies on an ordinary least squares (OLS) regression that weighs the SE and the PC data in proportion to the number of questions asked (Sawtooth Software 2002, Version 5).²⁶ We label the current version “weighted hybrid” estimation and denote it by the acronym WHSE (the SE suffix indicates reliance on SE data). Sawtooth Software also incorporates the SE responses in their Hierarchical Bayes estimation procedure by using the SEs to constrain the estimates of partworths both within a feature and between features to satisfy the ordinal conditions imposed by the SE data. For example, if a respondent’s responses to the SE questions indicate that picture quality is more important than battery life, then the Hierarchical Bayes parameters are restricted to satisfying this condition. We denote this algorithm with the acronym HBSE to indicate that the SE responses play a larger role in the estimation.

We also create a polyhedral hybrid by extending AC estimation to incorporate SE responses. To do so, we introduce constraints on the feasible polyhedron similar to those used by HBSE. For example, we impose a condition that picture quality is more important than battery life by using an inequality constraint on the polyhedron to exclude points in the partworth space that breach this condition. When the polyhedron becomes empty, we extend OPT4 to incorporate both the PC and SE constraints. We distinguish this method from the Analytic Center method by adding a suffix to the acronym: ACSE.

To compare WHSE, HBSE, and ACSE to their purebred progenitors, we must consider the accuracy of the SE data. If the SE data are perfectly accurate, then a model based on SEs alone will predict perfectly and the hybrids would be almost as accurate. On the other hand, if the SEs are extremely noisy, then the hybrids may actually predict worse than methods that do

²⁶ Earlier versions of ACA either weighed the scales equally (Version 3) or selected weights to fit purchase-intention questions (Version 4).

not use SE data. To examine these questions, we undertook a second set of simulation experiments.

To simulate SE responses we assume that respondents' answers to SE questions are unbiased but imprecise. In particular, we simulate response error in the SE answers by adding to the vector of true partworths, \bar{u} , a vector of independent identically-distributed normal error terms with variance σ_{se}^2 . We simulate two levels of SE response error – low error relative to PC responses ($\sigma_{se} = 10$) and high error relative to PC responses ($\sigma_{se} = 70$), truncating the SEs to a 0-to-100 scale.²⁷ We expect that these benchmarks should bound empirical situations. Recall that in the first set of simulations we assumed no SE errors ($\sigma_{se} = 0$), but discretized the scale. For consistency, we also use a discrete scale when there are non-zero SE errors. Based on these SE errors, we redo the simulations for each level of PC response errors and heterogeneity.

We summarize the results with Table 4 for lower numbers of questions ($q = 8$), where we report the most accurate question-design/estimation methods for each level of response error and heterogeneity. For ease of comparison, the earlier results (from Table 1) are summarized in the column labeled “Initial Simulations.”

There are three results of interest. First, when the SEs are more accurate than the PCs, then the hybrids do well. In this situation, the PC question design method matters less: Polyhedral, Fixed, and Random hybrids are not significantly different in accuracy. Second, the insights obtained from Tables 1-3 for population-level versus individual-level estimation continue to hold: HB or HBSE do well in homogeneous domains while AC or WHSE do well in heterogeneous domains. Third, when the SEs are noisy relative to the PCs, then the hybrid methods do not do as well as the purebred methods. Indeed, we expect a crossover point at some intermediate level of relative accuracy.

Table 4 also highlights the emergence of WHSE in some domains.²⁸ Of the hybrids tested, WHSE is the only method that makes use of the interval-scale properties of the SEs. These metric properties appear to help when the “signal-to-noise ratio” is high (more variation in

²⁷ For high SE errors truncation is approximately 50%; for low SE errors truncation is approximately 0% and 11%, respectively, for low and high heterogeneity. This is consistent with our manipulation of high vs. low information content of the SEs.

²⁸ If WHSE were not available, Polyhedral ACSE is best for low response errors and Polyhedral HBSE is best for high response errors. These results suggest that there is room for future research on how best to incorporate SE data with AC estimation.

true partworths, less error in the SEs). This result suggests that other methods which use interval-scaled properties of the SEs should do well in these domains – a topic for further hybrid development (e.g., Ter Hofstede, Kim, and Wedel 2002).

In summary, as in biology, where genetically-diverse offspring often have traits superior to their purebred parents, heterosis in conjoint analysis improves predictive accuracy in some domains. Furthermore, Polyhedral question design remains promising in these domains and many of the insights from our earlier simulations still hold for hybrid methods. Finally, we expect and obtain analogous results for larger numbers of questions ($q = 16$). We could identify no additional insight beyond Tables 1-4.

Table 4
The Impact of Self-Explicated (SE) Questions ($q = 8$)

Heterogeneity	Response Errors	Initial Simulations (no SEs)	Relatively Accurate SEs	Relatively Noisy SEs
Homogeneous	Low error	Polyhedral HB Fixed HB	Polyhedral HBSE Fixed HBSE Random HBSE	Polyhedral HB Fixed HB
	High error	Polyhedral HB ACA HB	Polyhedral HBSE Fixed HBSE Random HBSE	Polyhedral HBSE
Heterogeneous	Low error	Polyhedral AC	Polyhedral WHSE Fixed WHSE Random WHSE	Polyhedral AC
	High error	Polyhedral AC	Polyhedral WHSE Fixed WHSE Random WHSE	Polyhedral AC

6. Empirical Application and Test of Polyhedral Methods

While tests of internal validity are common in the conjoint-analysis literature, tests of external validity at the individual level are rare.²⁹ A search of the literature revealed four studies that predict choices in the context of natural experiments and one study based on a lottery choice.

²⁹ Some researchers report aggregate predictions relative to observed market share. See Bucklin and Srinivasan (1991), Currim (1981), Green and Srinivasan (1978), Griffin and Hauser (1993), Hauser and Gaskin (1984), McFadden (2000), Page and Rosenbaum (1989), and Robinson (1980).

Wittink and Montgomery (1979), Srinivasan (1988), and Srinivasan and Park (1997) all use conjoint analysis to predict MBA job choice. Samples of 48, 45, and 96 student subjects, respectively, completed a conjoint questionnaire prior to accepting job offers. The methods were compared on their ability to predict actual job choices. First preference predictions ranged from 64% to 76% versus random-choice percentages of 26-36%. In another natural experiment, Wright and Kriewall (1980) used conjoint analysis (Linmap) to predict college applications by 120 families. They were able to correctly predict 20% of the applications when families were prompted to think seriously about the features measured in conjoint analysis; 15% when they were not. This converts to a 16% improvement relative to their null model. Leigh, MacKay and Summers (1984) allocated 122 undergraduate business majors randomly to twelve different conjoint tasks designed to measure partworths for five features. Respondents indicated their preferences for ten calculators offered in lottery. There were no significant differences among methods with first-preference predictions in the range of 26-41% and percentage improvements of 28%. The authors also compared the performance of estimates based solely on SE responses and observed similar performance to the conjoint methods.

In this section, we test polyhedral methods with an empirical test involving an innovative laptop-computer carrying bag. Our test differs from the natural experiment studies because it is based on a controlled experiment in which we chose pareto sets of product features. At the time of our study, the product was not yet on the market and so respondents had no prior experience with it. The bag includes a range of separable product features, such as the inclusion of a mobile-phone holder, side pockets, or a logo. We focused on nine product features, each with two levels, and included price as a tenth feature. Price is restricted to two levels (\$70 and \$100) – the extreme prices for the bags in both the internal and external validity tests. We estimated the partworths associated with prices between \$70 and \$100 by linearly interpolating. A more detailed description of the product features can be found on the website cited earlier in this paper.

Because ACA is the dominant industry method for adaptive question design, we chose a product category where we expected ACA to perform well – a category where separable product features would lead to moderately accurate SE responses. We anticipate that SE responses are more accurate in categories where customers make purchasing decisions about features separately by choosing from a menu of features. In contrast, we expect SE responses to be less accurate for products where the features are typically bundled together, so that customers have little

experience in evaluating the importance of the individual features. If Polyhedral question design and/or estimation does well in this category, then, based on Table 4, we expect it to do well in categories where SE responses are less accurate.

Research Design

Subjects were randomly assigned to one of the three conjoint question design methods: Polyhedral (2 cells), Fixed, or ACA. We omitted Random question design because the Fixed question design method dominates Random design in Tables 2 and 3. After completing the respective conjoint tasks, all of the respondents were presented with the same validation exercises. The internal validation exercise involved four holdout metric paired-comparison (PC) questions, which occurred immediately after the sixteen PC questions designed by the respective conjoint methods. The external validation exercise was the selection of a laptop computer bag from a choice set of five bags. This exercise occurred in the same session as the conjoint tasks and holdout questions, but was separated from these activities by a filler task designed to cleanse memory (see Table 5).

Table 5
Detailed Research Design

Row	Polyhedral 1	Fixed	Polyhedral 2	ACA
1			Self-explicated	Self-explicated
2	Polyhedral paired comparison	Fixed paired comparison	Polyhedral paired comparison	ACA paired comparison
3	Internal validity task	Internal validity task	Internal validity task	Internal validity task
4			Purchase intentions	Purchase intentions
5	Filler task	Filler task	Filler task	Filler task
6	External validity task	External validity task	External validity task	External validity task

Conjoint Tasks

Recall that ACA requires five sets of questions. Pretests confirmed that all of the features were acceptable to the target market, allowing us to skip the unacceptability task. This left four remaining tasks: ranking of levels within features, self-explicated (SE) questions, metric paired-comparison (PC) questions, and purchase intention (PI) questions. ACA uses the SE questions to select the PC questions, thus the SE questions in ACA must come first, followed by the PC questions and then the PI questions. To test ACA fairly, we adopted this question order for the ACA condition.

The Fixed and Polyhedral question design techniques do not require SE or PI questions. Because asking the SE questions first could create a question-order effect, we asked only PC questions (not the SE or PI questions) prior to the validation task in the Fixed condition. To investigate the question-order effect we included two polyhedral data collection procedures: one that matched the Fixed design (Polyhedral 1) and one that matched ACA (Polyhedral 2). In Polyhedral 1 only PC questions preceded the validation task, while in Polyhedral 2, all of the questions preceded the validation task. This enables us to (a) explore whether the SE questions affect the responses to the PC questions and (b) evaluate the hybrid estimation methods that combine data from PC and SE questions.³⁰

The complete research design, including the question order, is summarized in Table 5. Questions associated with the conjoint tasks are highlighted in green (Rows 1, 2 and 4), while the validation tasks are highlighted in yellow (Rows 3 and 6). The filler task is highlighted in blue (Row 5). In this design, Polyhedral 1 can be matched with Fixed; Polyhedral 2 can be matched with ACA.

Internal Validity Task: Holdout PC Questions

Each of the question design methods designed sixteen metric paired-comparison (PC) questions. Following these questions, respondents answered four holdout PC questions – a test used extensively in the literature. The holdout profiles were randomly selected from an independent efficient design of sixteen profiles and did not depend on prior answers by that respon-

³⁰ Although the SE responses are collected in the Polyhedral 2 condition, they are not used in Analytic Center estimation or Polyhedral question design. However, they do provide the opportunity to test hybrid estimation methods.

dent. There was no separation between the sixteen initial questions and the four holdout questions, so that respondents were not aware that the questions were serving a different role.

Filler Task

The filler task was designed to separate the conjoint tasks and the external validity task. It was hoped that this separation would mitigate any memory effects that might influence how accurately the information from the conjoint tasks predicted which bags respondents chose in the external validity tasks. The filler task was the same in all four experimental conditions and comprised a series of questions asking respondents about their satisfaction with the survey questions. There was no significant difference in the responses to the filler task across the four conditions.

External Validity Task: Final Bag Selection

Respondents were told that they had \$100 to spend and were asked to choose among five bags. The five bags shown to each respondent were drawn randomly from an orthogonal fractional factorial design of sixteen bags. This design was the same across all four experimental conditions, so that there was no difference, on average, in the bags shown to respondents in each condition. The five bags were also independent of responses to the earlier conjoint questions. The price of the bags varied between \$70 and \$100 reflecting the difference in the anticipated market price of the features included with each bag. By pricing the bags in this manner we ensured that the choice set represented a Pareto frontier, as recommended by Elrod, Louviere, and Davey (1992), Green, Helsen and Shandler (1988), and Johnson, Meyer and Ghosh (1989).

Respondents were instructed that they would receive the bag that they chose. If the bag was priced at less than \$100, they were promised cash for the difference. In order to obtain a complete ranking, we told respondents that if one or more alternatives were unavailable, they might receive a lower ranked bag. The page used to solicit these rankings is presented in Figure

6.³¹ At the end of the study the chosen bags were distributed to respondents together with the cash difference (if any) between the price of the selected bag and \$100.

Self-Explicated and Purchase Intention Questions

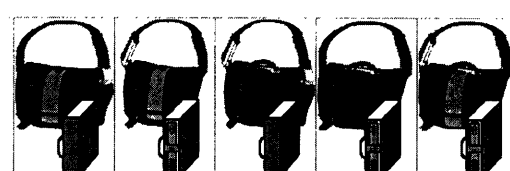
The self-explicated questions asked respondents to rate the importance of each of the ten product features. For a fair comparison to ACA, we used the wording for the questions, the (four-point) response scale, and the algorithm for profile selection proposed by Sawtooth Software (1996). For the purchase intention questions, respondents were shown six bags and we asked how likely they were to purchase each bag. We adopted the wording, response scale and algorithms for profile selection suggested by Sawtooth Software.

Figure 6
Respondents Choose and Keep a Laptop Computer Bag

Choose a bag to keep

The five available bags are illustrated below.
Please indicate your first, second, third, fourth and fifth choice

click on the
Features
for a reminder



Features	first	fifth	second	third	fourth
Size	Large	Medium	Medium	Large	Large
Color	Red/Gray	Red/Gray	Black	Black	Red/Gray
Logo	Yes	Yes	Yes	Yes	No
Handle	No	No	Yes	Yes	Yes
PDA	No	No	No	No	No
Cell Phone	No	Yes	Yes	No	Yes
Mesh Pocket	Yes	No	Yes	No	No
Sleeve Closure	Full Flap	Tab Velcro	Full Flap	Tab Velcro	Tab Velcro
Boot	Yes	Yes	Yes	Yes	No
Price	\$91	\$79	\$97	\$87	\$80

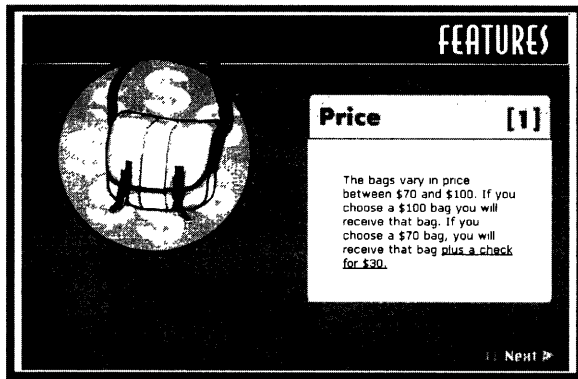
³¹ We acknowledge two tradeoffs in this design. The first is an endowment effect because we endow each respondent with \$100. The second is the lack of a “no bag” option. While both are interesting research opportunities and quite relevant to market forecasting, a priori neither should favor one of the three methods relative to the other; we expect no interaction between the endowment/forced-choice design and PC question design and leave such investigations to future research. However, the forced choice design might add noise to the most-accurate method relative to less-accurate methods. This would make it more difficult to achieve significant differences and is, thus, conservative. Pragmatically, we designed the task to maximize the power of the statistical comparisons of the four treatments. The forced-choice also helped to reduce the (substantial) cost of this research.

Subjects

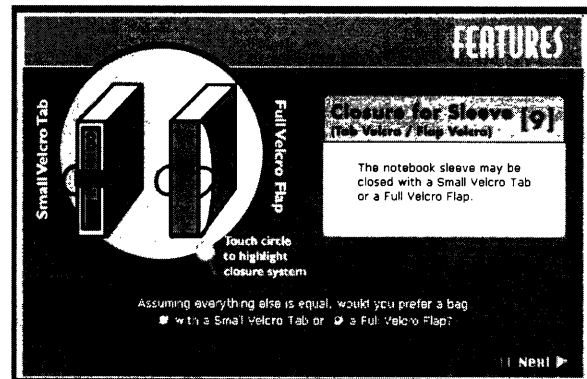
The subjects (respondents) were first-year MBA students. They were not informed about the objectives of the study, nor had they taken a course in which conjoint analysis was taught in detail. We received 330 complete responses (there was one incomplete response) from an e-mail invitation to 360 students – a response rate of over 91%. Pure random assignment (without quotas) yielded 80 subjects for the ACA condition, 88 for the Fixed condition, and 162 for the Polyhedral conditions broken out as 88 for the standard question order (Polyhedral 1) and 74 for the alternative question order (Polyhedral 2).

The questionnaires were pretested on a total of 69 subjects drawn from professional market research and consulting firms, former students, graduate students in Operations Research, and second-year students in an advanced marketing course that studied conjoint analysis. The pretests were valuable for fine-tuning the question wording and the web-based interfaces. By the end of the pretest, respondents found the questions unambiguous and easy to answer. Following standard scientific procedures, the pretest data were not merged with the experimental data. However, analysis of this small sample suggests that the findings agree directionally with those reported here, albeit not at the same level of significance.

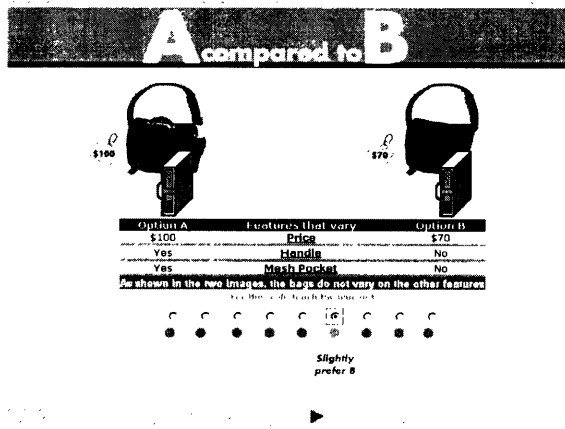
Figure 7
Example Screens from Questionnaires



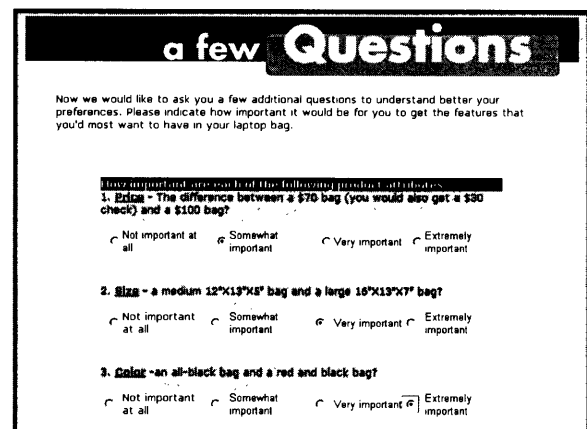
(a) Price as change from \$100



(b) Introduction of "sleeve" feature



(c) Metric paired-comparison (PC) question



(d) Self-explicated (SE) questions

Additional Details

Figure 7 illustrates some of the key screens in the conjoint analysis questionnaires. In Figure 7a respondents are introduced to the price feature. Figure 7b illustrates one of the dichotomous features – the closure on the sleeve. This is an animated screen that provides more detail as respondents move their pointing devices past the picture. Figure 7c illustrates one of the PC tasks. Respondents were asked to rate their relative preference for two profiles that varied on three features. Both text and pictures were used to describe the profiles. In the pictures, features that did not vary between the products were chosen to coincide with the respondent’s preferences for feature levels obtained in the tasks such as Figure 7b. The format was identical

for all four experimental treatments. Finally, Figure 7d illustrates the first three self-explicated questions. The full questionnaires for each treatment are available on the website cited earlier in this paper. We note that some of these website improvements (e.g., dynamically changing pictures) are not standard in Sawtooth Software’s implementation, thus, our tests should be considered a test of ACA question design (and estimation) rather than a test of Sawtooth Software’s commercial implementation.

7. Results of the Field Test

To evaluate the conjoint methods we calculated the Spearman rank-order correlation between the actual and observed rankings for the five bags shown to each respondent.³² We report the results in Table 6 using the same benchmark methods that we used for the Monte Carlo simulations.

Table 6
External Validity Tests: Correlation with Actual Choice
(Larger numbers indicate better performance)

	After 8 Questions		After 16 Questions	
Methods without SE data	Fixed Questions	Polyhedral 1 Questions	Fixed Questions	Polyhedral 1 Questions
Analytic Center (AC)	0.51	0.59	0.62	0.68
Hierarchical Bayes (HB)	0.53	0.57	0.61	0.64
Sample size	88	88	88	88
Methods that use SE data	ACA Questions	Polyhedral 2 Questions	ACA Questions	Polyhedral 2 Questions
WHSE Estimation	0.66	0.68	0.68	0.72
ACSE Estimation	0.63	0.70	0.65	0.71
HBSE Estimation	0.64	0.73	0.65	0.74
Sample size	80	74	80	74

³² As an alternative metric, we compared how well the methods predicted which product the respondents favored. The two metrics provide a similar pattern of results and so, for ease of exposition, we focus on the correlation measure. There are additional reasons to focus on correlations. First-choice prediction is a dichotomous variable highly dependent upon the number of items in the choice set. In addition, it provides less power because it has higher variance than the Spearman correlation, which is based on a rank order of five items.

In the simulation analysis we had the luxury of large sample sizes (500 respondents) and we were able to completely control for respondent heterogeneity. Although the sample sizes in Table 6 are large compared to previous tests of this type, they are small compared to the simulation analysis. As a result, none of the differences across methods are significant at the 0.05 level in independent-sample t-tests. However, these independent-sample t-tests do not use all of the information available in the data. We also evaluate significance by an alternative method that pools the correlation measures calculated after each additional PC question. This results in a total of sixteen observations for each respondent.

To control for heteroscedasticity we estimate a separate intercept for each question number. We also controlled for respondent heterogeneity in the randomly-assigned conditions with a null model that assumes that the ten laptop bag features are equally important. If, despite the random assignment of respondents to conditions, the responses in one condition are more consistent with the null model, then the comparisons would be biased in favor of this condition. We control for such potential heterogeneity by including a measure describing how accurately the equal-weights (null) model performs on the predictive correlations. The complete specification for this model is described in Equation 1, where r indexes the respondents and q indexes the number of PC questions used in the partworth estimates. The α 's and β 's are coefficients in the regression and ε_{rq} is an error term.

$$(1) \quad Correlation_{rq} = \sum_{q=1}^{16} \alpha_q Question_q + \sum_{m=1}^{M-1} \beta_m Method_m + \gamma EqualWeights_r + \varepsilon_{rq}$$

The *Question* and *Method* terms refer to dummy variables identifying the question and method effects. The *EqualWeight* variable measures the correlation obtained for respondent r between the actual rankings and the rankings obtained from an equal weights model. Under this specification, the β coefficients represent the expected increase or decrease in this correlation across questions due to Method m relative to an arbitrarily chosen base method. Positive (negative) values for the β coefficients indicate that the correlations between the actual and predicted rankings are higher (lower) for Method m than the base method.

We further control for potential heteroscedasticity introduced by the panel nature of the data by calculating robust standard errors (White 1980). We also estimated a random effects model, but there were almost no difference in the coefficients of interest. Moreover, the Haus-

man specification test favored the fixed-effects specification. The findings are summarized in Table 7.

Table 7
External Validity Tests: Conclusions from the Multivariate Analysis

	Without SE Questions	With SE Questions
Comparison of Estimation Methods		
Fixed Questions	HB > AC	
Polyhedral 1 Questions	AC >> HB	
ACA Questions		WHSE > HBSE > ACSE
Polyhedral 2 Questions		HBSE >>> WHSE > ACSE
Comparison of Question design Methods		
AC Estimation	Polyhedral 1 >>> Fixed	
HB Estimation	Polyhedral 1 >>> Fixed	
WHSE Estimation		Polyhedral 2 > ACA
ACSE Estimation		Polyhedral 2 >> ACA
HBSE Estimation		Polyhedral 2 >>> ACA

Method m > Method n: Method *m* is more accurate than Method *n* but the difference is not significant.
Method m >> Method n: Method *m* is significantly more accurate than Method *n* ($p < 0.05$).
Method m >>> Method n: Method *m* is significantly more accurate than Method *n* ($p < 0.01$).

Comparison of Estimation Methods

We compare the accuracy of the different estimation methods by comparing the findings in Table 6 within a column (for a specific set of questions) and looking to Table 7 for significance tests. This comparison holds the question design constant and varies the estimation method. For those experimental cells that were designed to obtain estimates without the SE questions, Hierarchical Bayes and Analytic Center estimation offer similar predictive accuracy for Fixed questions, but Analytic Center estimation performs better for Polyhedral questions.

If SE responses are available, the preferred estimation method appears to depend upon both the question design method and the number of PC responses used in the estimation. For Polyhedral questions HBSE performs extremely well for low numbers of PC questions, perhaps due to its use of population level data. However, increasing the number of PC responses yields less improvement in the accuracy of HBSE relative to WHSE. After sixteen questions all three

estimation methods converge to comparable accuracy levels, suggesting that there are sufficient data at the individual level to provide estimates that need not depend on population distributions. When using PC questions designed by ACA, WHSE out-performs HBSE, albeit not significantly so.

Comparison of Question Design Methods

The findings in Table 6 also facilitate comparison of the question design methods. Comparing across columns (within rows) in Table 6 holds the estimation method constant and varies the question design. The findings favor the two conditions in which the Polyhedral question design was used. When the SE measures were not collected, the Polyhedral question design yielded significantly ($p < 0.01$) more accurate predictions than the Fixed design. This holds true irrespective of the estimation method.

When SE responses were collected, the Polyhedral question design was more accurate than ACA across every estimation method, although the difference was not significant for WHSE. Detailed investigation reveals that for every estimation method we tested, the estimates derived using the Polyhedral questions outperform the corresponding estimates derived using ACA questions after each and every question number.

The Incremental Predictive Value of the SE Questions and of PC Questions

In this category, Table 6 suggests that hybrid methods that use both SE and PC questions consistently outperform methods that rely on PC questions alone. Thus, in this category, the SE questions provide incremental predictive ability. We caution that the product category was chosen at least in part because the SE responses were expected to be accurate. The simulations suggest that this improvement in accuracy may not be true in all domains.

We also evaluate whether the PC responses contributed incremental accuracy. Predictions that use the SE responses alone (without the PC responses) yield an average correlation with actual choice of 0.64. This is lower than the performance of the best methods that use both SE and PC responses and comparable at $q = 16$ to those methods that do not use SE responses (see Table 6). We conclude (in this category) that the sixteen PC questions provide roughly the same amount of information as the ten SE questions and that, for methods that use both, the PC data add incremental predictive ability. This conclusion is consistent with previous evidence in

the literature (Green, Goldberg, and Montemayor 1981; Huber, et. al. 1993, Johnson 1999; Leigh, MacKay, and Summers 1984).

The Internal Validity Task

We repeated the analysis of question design and estimation methods using the correlation measures from the internal validity (holdout questions) task. Details are in Appendix 2. The results for internal validity are similar to the results for external validity. However, there are two differences worth noting. First, while HBSE predicted better than WHSE for Polyhedral question design in the choice task, there was no significant difference in the holdout task. Second, while Polyhedral question design was significantly better than Fixed design for the choice task, there were no significant differences for the holdout task.

Question Order Effects: Polyhedral 1 versus Polyhedral 2

Polyhedral 1 and Polyhedral 2 varied in question order; the SE questions preceded the PC questions in Polyhedral 2 but not in Polyhedral 1. Otherwise, both methods used the same question design algorithm – an algorithm that does not use SE data. Nonetheless, question order might influence the accuracy of the PC responses. If the SE questions “wear out” or tire respondents, causing them to pay less attention to the PC questions, we might expect that inclusion of the SE questions will degrade the accuracy of the PC responses. Alternatively, the SE questions may improve the accuracy of the PC questions by acting as a training or “warm-up” task which helps respondents clarify their values, increasing the accuracy of the PC questions (Green, Krieger and Agarwal 1991; Huber, et. al. 1993; Johnson 1991).

By comparing the two experimental cells we investigate whether the prior SE questions affected the accuracy of the respondents’ PC responses. The predictive accuracy of the two conditions are not statistically different ($t = -0.05$ for AC estimation, the preferred estimation method from Tables 1 and 7). This suggests that by the sixteenth question any wear out or warm-up/learning had disappeared. However, there might still be an effect for the early questions. When we estimate performance of AC estimation using a version of Equation 1, the effect is not significant for external validity task ($t = 0.68$), but is significant for the internal validity task ($t = 2.60$). In summary, the evidence is mixed. There is no evidence that the SE questions improve or degrade the accuracy of the PC questions for the choice task, but they might improve accuracy for the hold out task. Further testing is warranted. For example, if the first cut of the polyhedron

is critical, then it is important that the first PC question be answered as accurately as feasible. The use of SE questions might sensitize the respondent and enhance the accuracy of the first PC question. Alternatively, researchers might investigate other warm-up questions, such as those in which the respondent configures an ideal product (Dahan and Hauser 2002).

Summary of the Field Test

In the field test, Polyhedral question design appears to be the most accurate of the tested question design methods. When SE data are available, the most accurate estimation methods were the HBSE and WHSE hybrids. If SE data were unavailable, the most accurate estimation method was AC for Polyhedral questions and HB for Fixed questions.

To compare the field test to the simulations, we must identify the relevant domain. Fortunately, estimates of heterogeneity and PC response errors are a by-product of the Hierarchical Bayes estimation and we can use HB to estimate SE errors. These estimates suggest high levels of heterogeneity ($\sigma_u^2 \approx 29$) and PC response errors ($\sigma_{pc}^2 \approx 43$), but moderately low SE response errors ($\sigma_{pc}^2 \approx 18$). When SEs are unavailable, the simulations predict that in this domain: (1) Polyhedral question design should be better than Fixed for both estimation methods, (2) AC should be much better than HB for Polyhedral questions, (3) AC should remain better than HB for Fixed questions, but the difference is not as large. The significant findings in Table 7 are consistent with (1) and (2). Contrary to (3), HB is better for Fixed questions, but not significantly so.

For accurate SEs, the simulations predict that in this domain: (4) Polyhedral questions will remain strong for hybrid estimation methods, but the differences among question design methods will be less for hybrid methods than for purebred methods, (5) hybrid estimation methods will outperform the purebred methods, and (6) WHSE will outperform ACSE (in the detailed simulation data for this domain HBSE also outperforms ACSE). Predictions (4), (5), and (6) hold true in the field test.

It is always difficult to compare field data to simulations because, despite experimental controls, there may be unobserved phenomena in the field test that are not captured in the simulations. However, the two type of data are remarkably consistent, albeit not perfectly so.

8. Product Launch

Subsequent to our research, Timbuk2 launched the laptop bags with features similar to those tested including multiple sizes, custom colors, logo options, accessory holders (PDA and cellular phone), mesh pockets, and laptop sleeves. Timbuk2 considers the product a success – it is selling well and is profitable. We now compare the laboratory experiment and the national launch. However, we do so with caution because the goal of the field test was to compare methods rather than to forecast the national launch. By design we used a student sample rather than a national sample, offered only two color combinations, and did not offer the large size bag. Furthermore, one tested feature, the “boot,” was not included in the national launch because production cost (and feasibility) exceeded the price that could be justified. One feature, a bicycle strap, was added based on managerial judgment.

There were five comparable features that appeared in both the field test and the national launch. With the above caveats in mind, the correlation of the predicted feature shares from the conjoint analyses with those observed in the marketplace was 0.9, which was significant. (By feature share we mean percent of customers who chose each of the five features.) Predictions with various null models were not significant. Unfortunately, these data do not provide sufficient power to compare the relative accuracies of the methods nor report correlations to more than one significant figure.

9. Conclusions and Future Research

Recent developments in math programming provide new methods for designing metric paired-comparison questions and estimating partworths. The question design method uses a multidimensional polyhedron to characterize feasible parameters and focus questions to reduce the size of the polyhedron as fast as possible. The estimation method uses the analytic center to approximate the center of the polyhedron. This centrality estimate summarizes what is known about the feasible set of parameters. Our goals in this paper were to (1) propose practical algorithms using polyhedral methods, (2) demonstrate their feasibility, (3) test their potential in a variety of domains, (4) compare their theoretical accuracy relative to existing methods, and (5) compare their predictive accuracy relative to existing methods in a realistic empirical situation. The field test was designed to match one of the theoretical domains that was favorable to existing

methods. The overall conclusion is that polyhedral methods are worth further development, experimentation, and study. Detailed findings are summarized in Table 8.

Table 8
Detailed Summary of Findings

Feasibility

- The Polyhedral question-design method can design questions in real time for a realistic number of parameters.
- The Analytic Center estimation heuristic yields real time partworth estimates that provide reasonable accuracy with little or no bias.

Monte Carlo Experiments

- Polyhedral question design shows the most promise for lower numbers of questions where it does well in all tested heterogeneity and response-error domains.
- Fixed question design remains best for larger numbers of questions, but Poly/Random does well for homogeneous populations and high response errors.
- AC estimation shows promise for heterogeneous populations and low response errors; HB performs well when the population is homogeneous and response errors are high.
- When self-explicated (SE) data are available and relatively noisy:
 - Polyhedral question design continues to perform well.
 - Purebred estimation methods may be preferred.
- When SE data are available and relatively accurate:
 - Question design is relatively less important.
 - For homogeneous populations, HBSE is the best estimation method.
 - For heterogeneous populations, WHSE is the best estimation method.

External-Validity Experiment

- The field test domain had high heterogeneity and PC response errors, and moderately low SE response errors. The findings are consistent with the Monte Carlo results in this domain.
- Polyhedral question design shows promise irrespective of the availability of SE data. This is true for all tested estimation methods.
- When SE data are unavailable, Analytic Center estimation is a viable alternative to Hierarchical Bayes estimation, especially when paired with Polyhedral question design.
- When SE data are available, existing hybrid estimation methods (HBSE and WHSE) appear to outperform the ACSE hybrid.

Product Launch

- The laptop bags were launched to the market, but we must be cautious when evaluating predictive accuracy because there were many differences between the empirical experiment and the market launch. In addition, there were insufficient data for relative comparisons.
 - With these caveats, the conjoint analyses correlate well with the marketplace outcome.
-

We are encouraged by the performance of polyhedral methods in the Monte Carlo and external-validity experiments. We feel that with further research new algorithms based on polyhedral methods have the potential to become easier to use, more accurate, and applicable in a broader set of domains. We close by highlighting some of the opportunities.

Theoretical Improvements

Error theory. Our proposed heuristic (OPT4) provides a practical means by which to use the Analytic Center to obtain partworth estimates from noisy data. Although the maximum error, $\bar{\delta}$, is likely to increase with the number of questions, the mean error, $\bar{\delta}/q$, is likely to decrease. Furthermore, the Analytic Center estimates appear to be unbiased, except when there is endogeneity in question design. However, we have not yet developed a formal proof of either hypothesis. More generally, it might be possible to derive the error-handling heuristics from more-fundamental distributional assumptions.

Algorithmic improvements for question design. There are a number of ways in which Polyhedral question design might be improved. For example, the proposed algorithm reverts to random selection when the polyhedron becomes empty (around $q = 8$ when $p = 10$). We might explore algorithms that use relaxed constraints (OPT4) after the initial polyhedron becomes empty. Multi-step look-ahead algorithms might yield further improvements.

Improved hybrid estimation. Analytic Center estimation appears to do quite well when self-explicated (SE) data are not available or when SE data are relatively noisy. However, when the SE data are relatively accurate, existing methods are better. In particular, for some domains a method that directly exploits the metric properties of the SEs seems to be best. We proposed ACSE as a natural way to mix AC estimation with SE constraints, but there might be other methods that make better use of the metric properties of the SEs. Table 4 also suggests that SE metric properties might be useful with Hierarchical Bayes hybrids, such as the Ter Hofstede, Kim, and Wedel (2002) algorithm.

Complexity constraints. Evgeniou, Boussios, and Zacharia (2002) demonstrate that “Support Vector Machines” (SVM) can improve estimation by automatically balancing complexity of the partworth specification with fit. These researchers are now exploring a hybrid between Analytic-Center and SVM estimation for stated-choice data – an exciting development that can

deal with non-linearities in polyhedral specifications. We might also extend the algorithm in Appendix 1 to incorporate complexity considerations directly either by using relaxed constraints (thick hyperplanes as in Figure 2b) or the direct use of OPT4 in the definition of the hyperplanes.

Monte Carlo Experiments

Addressing the “why.” Our heuristics are designed to obtain better questions by reducing the feasible polyhedron as rapidly as possible. In an analogy to the theory of efficient questions, e.g., D-efficiency, this heuristic focuses the next question on that portion of the parameter space where we have the least information. Future Monte Carlo experiments might test this hypothesis or identify an alternative explanation. Such theory might lead to further improvements in the algorithm.

Learning and wear-out. The comparison of question order (Polyhedral 1 vs. Polyhedral 2) suggests that the SE questions might increase the accuracy of the paired-comparison questions by acting as warm-up questions. This is an example of the more general issue that response errors might depend upon q . For example, exploratory simulations, using $\sigma_{pc} = \kappa_1 + \kappa_2 q$, suggest that learning (positive κ_2) causes mean absolute error (MAE) to decline more rapidly as the number of questions, q , increases. Wear out (negative κ_2) yields a U-shaped function. Other simulations might explore further many behavioral issues affecting response accuracy (Tourangeau, Rips and Rasinski 2000).

Interactions. Analytic Center estimation is a by-product of Polyhedral question design. Tables 1, 6, and 7 suggest that there might be interactions among question design and estimation methods in terms of MAE or predictive ability. Although none of these interactions were significant in our data, interactions are worth further exploration. For example, while AC does well when the SE data are unavailable or relatively noisy, ACSE does not do as well for relatively accurate SE data.

Other domains. Our simulations explore heterogeneity, response error, and the number of questions. We made a number of decisions such as the manner in which we specified heterogeneity (normally distributed with mean at the center of the range) and response error (normally distributed with no bias). Initial simulations suggested that the results were not sensitive to these assumptions, but more systematic exploration might identify interesting phenomena that we were not able to explore. Simulations for extremely large p or q might also yield new insight.

Other criteria. We chose MAE as our evaluative criteria. Exploratory simulation suggested that root mean squared error (RMSE) appeared to be proportional to MAE. We might also explore other criteria such as predictive accuracy (analogous to Table 6), the dollar value of product features (Jedidi, Jagpal and Manchanda 2003; Ofek and Srinivasan 2002), or the incremental explanatory power of each question and question type.

Empirical Tests in Other Categories

Known applications. Polyhedral methods are beginning to diffuse. Sawtooth Software, Inc. now offers a polyhedral option to its ACA software, Harris Interactive, Inc. has begun initial testing, and National Family Opinion, Inc. is exploring feasibility. Sawtooth Software has completed an empirical test of internal validity using a Poly/ACA question design algorithm (Orme and King 2002). In their data, on average, the ACA portion chose 63% of the paired-comparison questions. They observed no significant differences between the methods after $q = 30$. We have not been able to obtain for their application estimates of heterogeneity, PC response error, SE response error, or performance for low q .

Other product categories. We choose a category with separable features. In this category, the SE data were relatively accurate and, thus, the ACA benchmark was not handicapped. This domain matched one of the $3 \times 2 \times 2$ classes of categories in Table 4 and the empirical data were consistent with the Monte Carlo data. We hypothesize that the Monte Carlo experiments are consistent with the eleven other classes of categories, but further empirical tests might explore this hypothesis further.

References

- Allenby, Greg M. and Peter E. Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, (March/April), 57-78.
- Arora, Neeraj, Greg M. Allenby and James L. Ginter (1998), "A Hierarchical Bayes Model of Primary and Secondary Demand," *Marketing Science*, 17, 1, 29-44.
- and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28, (September), 273-283.
- Bucklin, Randolph E. and V. Srinivasan (1991), "Determining Interbrand Substitutability Through Survey Measurement of Consumer Preference Structures," *Journal of Marketing Research*, 28, (February), 58-71.
- Currim, Imran S. (1981), "Using Segmentation Approaches for Better Prediction and Understanding from Consumer Mode Choice Models," *Journal of Marketing Research*, 18, (August), 301-309.
- Dahan, Ely and John R. Hauser (2002), "The Virtual Customer," *Journal of Product Innovation Management*, 19, 5, (September), 332-354.
- Dawes, Robin M. and Bernard Corrigan (1974), "Linear Models in Decision Making," *Psychological Bulletin*, 81, (March), 95-106.
- Einhorn, Hillel J. (1971), "Use of Nonlinear, Noncompensatory, Models as a Function of Task and Amount of Information," *Organizational Behavior and Human Performance*, 6, 1-27,
- 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.
- Evgeniou, Theodoros, Constantinos Boussios, and Giorgos Zacharia (2002), "Generalized Robust Conjoint Estimation," Working Paper, (Fontainebleau, France: INSEAD).
- Freund, Robert (1993), "Projective Transformations for Interior-Point Algorithms, and a Super-linearly Convergent Algorithm for the W-Center Problem," *Mathematical Programming*, 58, 385-414.
- , Robert, R. Roundy, and M.J. Todd (1985), "Identifying the Set of Always-Active Constraints in a System of Linear Inequalities by a Single Linear Program," WP 1674-85, MIT Sloan School of Management.

- Green, Paul E., (1984), "Hybrid Models for Conjoint Analysis: An Expository Review," *Journal of Marketing Research*, pp. 155-169.
- , Kristiaan Helsen and Bruce Shandler (1988), "Conjoint Internal Validity Under Alternative Profile Presentations," *Journal of Consumer Research*, 15, (December), 392-397.
- , Stephen M. Goldberg, and Mila Montemayor (1981), "A Hybrid Utility Estimation Model for Conjoint Analysis," *Journal of Marketing*, 33-41.
- , Abba Krieger, and Manoj K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research*, 23, 2, (May), 215-222.
- , ----, and Jerry Wind (2001), "Thirty Years of Conjoint Analysis: Reflections and Prospects," *Interfaces*, 21, 3, Part 2 of 2, (May-June), S56-S73.
- , ----. and ---- (2002), "Buyer Choice Simulators, Optimizers, and Dynamic Models," *Advances in Marketing Research: Progress and Prospects*, (Philadelphia, PA: Wharton Press).
- and V. Srinivasan (1978), "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5, 2, (September), 103-123.
- and ---- (1990), "Conjoint Analysis in Marketing: New Developments With Implications for Research and Practice," *Journal of Marketing*, 54, 4, (October), 3-19.
- and ---- (1978), "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5, 2, (September), 103-123.
- Griffin, Abbie and John R. Hauser (1993), "The Voice of the Customer," *Marketing Science*, vol. 12, No. 1, (Winter), 1-27.
- Gritzmann P. and V. Klee (1993), "Computational Complexity of Inner and Outer J-Radii of Polytopes in Finite-Dimensional Normed Spaces," *Mathematical Programming*, 59(2), pp. 163-213.
- Hadley, G. (1961), *Linear Algebra*, (Reading, MA: Addison-Wesley Publishing Company, Inc.)
- Hauser, John R., and Steven P. Gaskin (1984), "Application of the 'Defender' Consumer Model," *Marketing Science*, Vol. 3, No. 4, (Fall), 327-351.
- and Vithala Rao (2003), "Conjoint Analysis, Related Modeling, and Applications," *Advances in Marketing Research: Progress and Prospects*, Jerry Wind, Ed., forthcoming.
- and Steven M. Shugan (1980), "Intensity Measures of Consumer Preference," *Operation Research*, Vol. 28, No. 2, (March-April), 278-320.
- Hoaglin, David C., Frederick Mosteller, and John W. Tukey (1983), *Understanding Robust and Exploratory Data Analysis*, (New York, NY: John Wiley & Sons, Inc.).

- Huber, Joel (1975), "Predicting Preferences on Experimental bundles of Attributes: A Comparison of Models," *Journal of Marketing Research*, 12, (August), 290-297.
- , Dick R. Wittink, John A. Fiedler, and Richard Miller (1993), "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice," *Journal of Marketing Research*, 105-114.
- and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33, (August), 307-317.
- Jedidi, Kamel; Puneet Manchanda, and Sharan Jagpal (2003), "Measuring Heterogeneous Reservation Prices For Product Bundles," *Marketing Science*, 22, 1, (Winter), forthcoming.
- Johnson, Eric J., Robert J. Meyer, and Sanjoy Ghose (1989), "When Choice Models Fail: Compensatory Models in Negatively Correlated Environments," *Journal of Marketing Research*, 255-270.
- Johnson, Richard (1987), "Accuracy of Utility Estimation in ACA," Working Paper, Sawtooth Software, Sequim, WA, (April).
- (1991), "Comment on `Adaptive Conjoint Analysis: Some Caveats and Suggestions,'" *Journal of Marketing Research*, 28, (May), 223-225.
- (1999), "The Joys and Sorrows of Implementing HB Methods for Conjoint Analysis," Working Paper, Sawtooth Software, Sequim, WA, (November).
- Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. C. Lee (1985), *The Theory and Practice of Econometrics*, (New York, NY: John Wiley and Sons).
- Karmarkar, N. (1984), "A New Polynomial Time Algorithm for Linear Programming," *Combinatorica*, 4, 373-395.
- Kuhfeld, Warren F. , Randall D. Tobias, and Mark Garratt (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, 31, 4, (November), 545-557.
- 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, 4, (November), 456-462.
- 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, 173-191.
- Liechty, John, Venkatram Ramaswamy, Steven Cohen (2001), "Choice-Menus for Mass Customization: An Experimental Approach for Analyzing Customer Demand With an Ap-

- plication to a Web-based Information Service,” *Journal of Marketing Research*, 38, 2, (May).
- Louviere, Jordan J., David A. Hensher, and Joffre D. Swait (2000), *Stated Choice Methods: Analysis and Application*, (New York, NY: Cambridge University Press).
- McFadden, Daniel (2000), “Disaggregate Behavioral Travel Demand’s RUM Side: A Thirty-Year Retrospective,” Working Paper, University of California, Berkeley, (July).
- Moore, William L. (2003), “A Cross-Validity Comparison of Conjoint Analysis and Choice Models,” Working Paper, (Salt Lake City, Utah: University of Utah), February.
- 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*, 261-274.
- Nesterov, Y. and A. Nemirovskii (1994), “Interior-Point Polynomial Algorithms in Convex Programming,” SIAM, Philadelphia.
- Ofek, Elie and V. Srinivasan (2002), “How Much Does the Market Value an Improvement in a Product Attribute?” *Marketing Science*, 21, 4, (Fall), 98–411.
- Orme, Bryan (1999), “ACA, CBC, of Both?: Effective Strategies for Conjoint Research,” Working Paper, Sawtooth Software, Sequim, WA.
- and W. Christopher King (2002), “Improving ACA Algorithms: Challenging a Twenty-year-old Approach,” *Presentation at Advanced Research Technology Conference*, June 3.
- Page, Albert L. and Harold F. Rosenbaum (1989), “Redesigning Product Lines with Conjoint Analysis: How Sunbeam Does It,” *Journal of Product Innovation Management*, 4, 120-137.
- Reibstein, David, John E. G. Bateson, and William Boulding (1988), “Conjoint Analysis Reliability: Empirical Findings,” *Marketing Science*, 7, 3, (Summer), 271-286.
- Robinson, P. J. (1980), “Applications of Conjoint Analysis to Pricing Problems,” in *Market Measurement and Analysis: Proceedings of the 1979 ORSA/TIMS Conference on Marketing*, David B. Montgomery and Dick Wittink, eds., (Cambridge, MA: Marketing Science Institute), 183-205.
- Sandor, Zsolt and Michel Wedel (2001), “Designing Conjoint Choice Experiments Using Managers’ Prior Beliefs,” *Journal of Marketing Research*, 38, 4, (November), 430-444.
- and ---- (2002), “Profile Construction in Experimental Choice Designs for Mixed Logit Models,” *Marketing Science*, 21, 4, (Fall), 398–411.
- Sawtooth Software, Inc. (1996), “ACA System: Adaptive Conjoint Analysis,” *ACA Manual*,

- (Sequim, WA: Sawtooth Software, Inc.)
- (2001), "The ACA/Hierarchical Bayes Technical Paper," (Sequim, WA: Sawtooth Software, Inc.)
- (2002), "ACA 5.0 Technical Paper," (Sequim, WA: Sawtooth Software, Inc.).
- Sonnevend, G. (1985a), "An 'Analytic' Center for Polyhedrons and New Classes of Global Algorithms for Linear (Smooth, Convex) Programming," *Proceedings of the 12th IFIP Conference on System Modeling and Optimization*, Budapest.
- (1985b), "A New Method for Solving a Set of Linear (Convex) Inequalities and its Applications for Identification and Optimization," Preprint, Department of Numerical Analysis, Institute of Mathematics, Eötvös University, Budapest, 1985.
- Srinivasan, V. (1988), "A Conjunctive-Compensatory Approach to The Self-Explication of Multiattributed Preferences," *Decision Sciences*, 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.
- Srinivasan, V. and Allan D. Shocker (1973), "Linear Programming Techniques for Multidimensional Analysis of Preferences," *Psychometrika*, 38, 3, (September), 337-369.
- Ter Hofstede, Frenkel, Youngchan Kim, and Michel Wedel (2002), "Bayesian Prediction in Hybrid Conjoint Analysis," *Journal of Marketing Research*, 39, (May), 253-261.
- Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski (2000), *The Psychology of Survey Response*, (New York, NY: Cambridge University Press), 197-229.
- Toubia, Olivier, Duncan Simester, John R. Hauser (2003), "Polyhedral Methods for Adaptive Choice-based Conjoint Analysis," working paper, Cambridge, MA: Center for Innovation in Product Development, MIT, (February).
- Tukey, John W. (1960), "A Survey of Sampling from Contaminated Distributions," I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, and H. Mann (Eds.) *Contributions to Probability and Statistics*, (Stanford, CA: Stanford University Press), 448-485.
- Urban, Glen L. and Gerald M. Katz, "Pre-Test Market Models: Validation and Managerial Implications," *Journal of Marketing Research*, Vol. 20 (August 1983), 221-34.
- Vaidja, P. (1989), "A Locally Well-Behaved Potential Function and a Simple Newton-Type Method for Finding the Center of a Polytope," in: N. Megiddo, ed., *Progress in Mathematical Programming: Interior Points and Related Methods*, Springer: New York, 79-90.

- White, H (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.
- Wittink, Dick R. and Philippe Cattin (1981), "Alternative Estimation Methods for Conjoint Analysis: A Monte Carlo Study," *Journal of Marketing Research*, 18, (February), 101-106.
- and David B. Montgomery (1979), "Predictive Validity of Trade-off Analysis for Alternative Segmentation Schemes," *1979 AMA Educators' Conference Proceedings*, Neil Beckwith ed., Chicago, IL: American Marketing Association.
- Wright, Peter and Mary Ann Kriewall (1980), "State-of-Mind Effects on Accuracy with which Utility Functions Predict Marketplace Utility," *Journal of Marketing Research*, 17, (August), 277-293.
- Yang, Sha; Greg M. Allenby, and Geraldine Fennell (2002), "Modeling Variation in Brand Preference: The Roles of Objective Environment and Motivating Conditions," *Marketing Science*, 21, 1, (Winter), 14-31.

Appendix 1: Mathematics of Fast Polyhedral Adaptive Conjoint Estimation

Consider the case of p parameters and q questions where $q \leq p$. Let u_j be the j^{th} parameter of the respondent's partworth function and let \vec{u} be the $p \times 1$ vector of parameters. Without loss of generality we assume binary features such that u_j is the high level of the j^{th} feature and constrain their values between 0 and 100. For more levels we simply recode the \vec{u} vector and impose constraints such as $u_m \leq u_h$. We handle such inequality constraints by adding slack variables, $v_{hm} \geq 0$, such that $u_h = u_m + v_{hm}$. Let r be the number of externally imposed constraints, of which $r' \leq r$ are inequality constraints.

Let $\vec{z}_{i\ell}$ be the $1 \times p$ vector describing the left-hand profile in the i^{th} paired-comparison question and let \vec{z}_{ir} be the $1 \times p$ vector describing the right-hand profile. The elements of these vectors are binary indicators taking on the values 0 or 1. Let X be the $q \times p$ matrix of $\vec{x}_i = \vec{z}_{i\ell} - \vec{z}_{ir}$ for $i = 1$ to q . Let a_i be the respondent's answer to the i^{th} question and let \vec{a} be the $q \times 1$ vector of answers for $i = 1$ to q . Then, if there were no errors, the respondent's answers imply $X\vec{u} = \vec{a}$. To handle additional constraints, we augment these equations such that X becomes a $(q+r) \times (p+r')$ matrix, \vec{a} becomes a $(q+r) \times 1$ vector, and \vec{u} becomes a $(p+r') \times 1$ vector. These augmented relationships form a polyhedron, $\mathbf{P} = \{ \vec{u} \in \mathcal{R}^{p+r'} \mid X\vec{u} = \vec{a}, \vec{u} \geq 0 \}$. We begin by assuming that \mathbf{P} is non-empty, that X is full-rank, and that no j exists such that $u_j = 0$ for all \vec{u} in \mathbf{P} . We later indicate how to handle these cases.

Finding an Interior Point of the Polyhedron

To begin the algorithm we first find a feasible interior point of \mathbf{P} by solving a linear program, LP1 (Freund, Roundy and Todd 1985). Let \vec{e} be a $(p+r') \times 1$ vector of 1's and let $\vec{0}$ be a $(p+r') \times 1$ vector of 0's; the y_j 's and θ are parameters of LP1 and \vec{y} is the $(p+r') \times 1$ vector of the y_j 's. (When clear in context, inequalities applied to vectors apply for each element.) LP1 is given by:

$$(LP1) \quad \max \sum_{j=1}^{p+r'} y_j, \quad \text{subject to: } X\vec{u} = \theta\vec{a}, \quad \theta \geq 1, \quad \vec{u} \geq \vec{y} \geq \vec{0}, \quad \vec{y} \leq \vec{e}$$

If $(\bar{u}^*, \bar{y}^*, \theta^*)$ solves LP1, then $\theta^{*-1} \bar{u}^*$ is an interior point of P whenever $\bar{y}^* > \bar{0}$. If there are some y_j 's equal to 0, then there are some j 's for which $u_j=0$ for all $\bar{u} \in P$. If LP1 is infeasible, then P is empty. We address these cases later in this appendix.

Finding the Analytic Center

The analytic center is the point in P that maximizes the geometric mean of the distances from the point to the faces of P . We find the analytic center by solving OPT1.

$$(OPT1) \quad \max \sum_{j=1}^{p+r'} \ln(u_j), \quad \text{subject to: } X\bar{u} = \bar{a}, \quad \bar{u} > \bar{0}$$

Freund (1993) proves with projective methods that a form of Newton's method will converge rapidly for OPT1. To implement Newton's method we begin with the feasible point from LP1 and improve it with a scalar, α , and a direction, \bar{d} , such that $\bar{u} + \alpha\bar{d}$ is close to the optimal solution of OPT1. (\bar{d} is a $(p+r') \times 1$ vector of d_j 's.) We then iterate subject to a stopping rule.

We first approximate the objective function with a quadratic expansion in the neighborhood of \bar{u} .

$$(A1) \quad \sum_{j=1}^{p+r'} \ln(u_j + d_j) \approx \sum_{j=1}^{p+r'} \ln(u_j) + \sum_{j=1}^{p+r'} \left(\frac{d_j}{u_j} - \frac{d_j^2}{2u_j^2} \right)$$

If we define U as a $(p+r') \times (p+r')$ diagonal matrix of the u_j 's, then the optimal direction solves OPT2:

$$(OPT2) \quad \max \bar{e}^T U^{-1} \bar{d} - (\gamma/2) \bar{d}^T U^{-2} \bar{d}, \quad \text{subject to: } X\bar{d} = \bar{0}$$

Newton's method solves OPT1 quickly by exploiting an analytic solution to OPT2. To see this, consider first the Karush-Kuhn-Tucker (KKT) conditions for OPT2. If \bar{z} is a $(p+r') \times 1$ vector parameter of the KKT conditions that is unconstrained in sign then the KKT conditions are written as:

$$(A2) \quad U^{-2} \bar{d} - U^{-1} \bar{e} = X^T \bar{z}$$

$$(A3) \quad X\bar{d} = \bar{0}$$

Multiplying A2 on the left by XU^2 , gives $X\bar{d} - XU\bar{e} = XU^2 X^T \bar{z}$. Applying A3 to this equation gives: $-XU\bar{e} = XU^2 X^T \bar{z}$. Since $U\bar{e} = \bar{u}$ and since $X\bar{u} = \bar{a}$, we have $-\bar{a} = XU^2 X^T \bar{z}$. Because

X is full rank and U is positive, we invert XU^2X^T to obtain $\vec{z} = -(XU^2X^T)^{-1} \vec{a}$. Now replace \vec{z} in A2 by this expression and multiply by U^2 to obtain $\vec{d} = \vec{u} - U^2X^T(XU^2X^T)^{-1} \vec{a}$.

According to Newton's method, the new estimate of the analytic center, \vec{u}' , is given by $\vec{u}' = \vec{u} + \alpha \vec{d} = U(\vec{e} + \alpha U^{-1} \vec{d})$. There are two cases for α . If $\|U^{-1} \vec{d}\| < 1/4$, then we use $\alpha=1$ because \vec{u} is already close to optimal and $\vec{e} + \alpha U^{-1} \vec{d} > \vec{0}$. Otherwise, we compute α with a line search.

Special Cases

If X is not full rank, XU^2X^T might not invert. We can either select questions such that X is full rank or we can make it so by removing redundant rows. Suppose that \vec{x}_k is a row of X such that $\vec{x}_k^T = \sum_{i=1, i \neq k}^{q+r} \beta_i \vec{x}_i^T$. Then if $a_k = \sum_{i=1, i \neq k}^{q+r} \beta_i a_i$, we remove \vec{x}_k . If $a_k \neq \sum_{i=1, i \neq k}^{q+r} \beta_i a_i$, then \mathbf{P} is empty and we employ OPT4 described later in this appendix.

If in LP1 we detect cases where some y_j 's = 0, then there are some j 's for which $u_j=0$ for all $\vec{u} \in \mathbf{P}$. In the later case, we can still find the analytic center of the remaining polyhedron by removing those j 's and setting $u_j = 0$ for those indices. If \mathbf{P} is empty we employ OPT4.

Finding the Ellipsoid and its Longest Axis

If \vec{u} is the analytic center and \bar{U} is the corresponding diagonal matrix, then Sonnevend (1985a, 1985b) demonstrates that $\mathbf{E} \subseteq \mathbf{P} \subseteq \mathbf{E}_{p+r'}$ where, $\mathbf{E} = \{ \vec{u} \mid X\vec{u} = \vec{a}, \sqrt{(\vec{u} - \vec{u})^T \bar{U}^{-2} (\vec{u} - \vec{u})} \leq 1 \}$ and $\mathbf{E}_{p+r'}$ is constructed proportional to \mathbf{E} by replacing 1 with $(p+r')$. Because we are interested only in the direction of the longest axis of the ellipsoids we can work with the simpler of the proportional ellipsoids, \mathbf{E} . Let $\vec{g} = \vec{u} - \vec{u}$, then the longest axis will be a solution to OPT3.

$$(OPT3) \quad \max \vec{g}^T \vec{g} \quad \text{subject to: } \vec{g}^T \bar{U}^{-2} \vec{g} \leq 1, \quad X\vec{g} = \vec{0}$$

OPT3 has an easy-to-compute solution based on the eigenstructure of a matrix. To see this we begin with the KKT conditions (where ϕ and γ are parameters of the conditions).

$$(A4) \quad \vec{g} = \phi \bar{U}^{-2} \vec{g} + X^T \vec{\gamma}$$

$$(A5) \quad \phi(\vec{g}^T \bar{U}^{-2} \vec{g} - 1) = 0$$

$$(A6) \quad \vec{g}^T \bar{U}^{-2} \vec{g} \leq 1, \quad X\vec{g} = \vec{0}, \quad \phi \geq 0$$

It is clear that $\bar{g}^T \bar{U}^{-2} \bar{g} = 1$ at optimal, else we could multiply \bar{g} by a scalar greater than 1 and still have \bar{g} feasible. It is likewise clear that ϕ is strictly positive, else we obtain a contradiction by left-multiplying A4 by \bar{g}^T and using $X\bar{g} = \bar{0}$ to obtain $\bar{g}^T \bar{g} = 0$ which contradicts $\bar{g}^T \bar{U}^{-2} \bar{g} = 1$. Thus, the solution to OPT3 must satisfy $\bar{g} = \phi \bar{U}^{-2} \bar{g} + X^T \bar{\gamma}$, $\bar{g}^T \bar{U}^{-2} \bar{g} = 1$, $X\bar{g} = \bar{0}$, and $\phi > 0$. We rewrite A4-A6 by letting I be the identity matrix and defining $\eta = 1/\phi$ and $\bar{\omega} = -\bar{\gamma}/\phi$.

$$(A7) \quad (\bar{U}^{-2} - \eta I) \bar{g} = X^T \bar{\omega}$$

$$(A8) \quad \bar{g}^T \bar{U}^{-2} \bar{g} = 1$$

$$(A9) \quad X\bar{g} = \bar{0}, \quad \phi > 0$$

We left-multiply A7 by X and use A9 to obtain $X\bar{U}^{-2} \bar{g} = XX^T \bar{\omega}$. Since X is full rank, XX^T is invertible and we obtain $\bar{\omega} = (XX^T)^{-1} X\bar{U}^{-2} \bar{g}$ which we substitute into A7 to obtain $(\bar{U}^{-2} - X^T (XX^T)^{-1} X\bar{U}^{-2}) \bar{g} = \eta \bar{g}$. Thus, the solution to OPT3 must be an eigenvector of the matrix, $M \equiv (\bar{U}^{-2} - X^T (XX^T)^{-1} X\bar{U}^{-2})$. To find out which eigenvector, we left-multiply A7 by \bar{g}^T and use A8 and A9 to obtain $\eta \bar{g}^T \bar{g} = 1$, or $\bar{g}^T \bar{g} = 1/\eta$ where $\eta > 0$. Thus, to solve OPT3 we maximize $1/\eta$ by selecting the smallest positive eigenvalue of M . The direction of the longest axis is then given by the associated eigenvector of M . We then choose the next question such that \bar{x}_{q+1} is most nearly collinear to this eigenvector subject any constraints imposed by the questionnaire design. (For example, in our simulation we require that the elements of \bar{x}_{q+1} be $-1, 0,$ or 1 .) The answer to \bar{x}_{q+1} defines a hyperplane orthogonal to \bar{x}_{q+1} .

We need only establish that the eigenvalues of M are real. To do this we recognize that $M = P\bar{U}^{-2}$ where $P = (I - X^T(XX^T)^{-1}X)$ is symmetric, i.e., $P = P^T$. Then if η is an eigenvalue of M , $\det(P\bar{U}^{-2} - \eta I) = 0$, which implies that $\det[\bar{U}(\bar{U}^{-1}P\bar{U}^{-1} - \eta I)\bar{U}^{-1}] = 0$. This implies that η is an eigenvalue of $\bar{U}^{-1}P\bar{U}^{-1}$, which is symmetric. Thus, η is real (Hadley 1961, 240).

Adjusting the Polyhedron so that it is non-Empty

P will remain non-empty as long as respondents' answers are consistent. However, in any real situation there is likely to be $q < p$ such that P is empty. To continue the polyhedral algorithm, we adjust P so that it is non-empty. We do this by replacing the equality constraint,

$X\bar{u} = \bar{a}$, with two inequality constraints, $X\bar{u} \leq \bar{a} + \bar{\delta}$ and $X\bar{u} \geq \bar{a} - \bar{\delta}$, where $\bar{\delta}$ is a $q \times 1$ vector of errors, δ_i , defined only for the question-answer imposed constraints. We solve the following optimization problem. Our current implementation uses the ∞ -norm where we minimize the maximum δ_i , but other norms are possible. The advantage of using the ∞ -norm is that (OPT 4) is solvable as a linear program.

$$\text{(OPT4)} \quad \min \|\bar{\delta}\| \quad \text{subject to: } X\bar{u} \leq \bar{a} + \bar{\delta}, \quad X\bar{u} \geq \bar{a} - \bar{\delta}, \quad \bar{u} \geq \bar{0},$$

At some point such that $q > p$, extant algorithms will outperform OPT4 and we can switch to those algorithms. Alternatively, a researcher might choose to switch to constrained regression (norm-2) or mean-absolute error (norm-1) when $q > p$. Other options include replacing some, but not all, of the equality constraints with inequality constraints. We leave these extensions to future research.

Appendix 2: Internal Validity Tests for Laptop Computer Bags

Table A2.1. Correlation with Actual Response

Methods without SE data	After 8 Questions		After 16 Questions	
	Fixed Questions	Polyhedral 1 Questions	Fixed Questions	Polyhedral 1 Questions
Analytic Center (AC)	0.65	0.69	0.80	0.79
Hierarchical Bayes (HB)	0.70	0.67	0.76	0.72
Sample size	88	87	88	87
Methods that use SE data	ACA Questions	Polyhedral 2 Questions	ACA Questions	Polyhedral 2 Questions
WHSE Estimation	0.77	0.81	0.81	0.84
ACSE Estimation	0.74	0.78	0.77	0.84
HBSE Estimation	0.76	0.80	0.78	0.82
Sample size	80	71	80	71

The missing observations reflect respondents who gave the same response for all four holdout questions (in which case the correlations were undefined).

Table A2.2. Conclusions from the Multivariate Analysis

	Without SE Questions	With SE Questions
Comparison of Estimation Methods		
Fixed Questions	HB > AC	
Polyhedral 1 Questions	AC >>> HB	
ACA Questions		WHSE > HBSE >>> ACSE
Polyhedral 2 Questions		WHSE > HBSE > ACSE
Comparison of Question design Methods		
AC Estimation	Polyhedral 1 > Fixed	
HB Estimation	Fixed > Polyhedral 1	
WHSE Estimation		Polyhedral 2 >>> ACA
ACSE Estimation		Polyhedral 2 >>> ACA
HBSE Estimation		Polyhedral 2 >>> ACA

Chapter 3: Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis

Abstract

Choice-based conjoint analysis (CBC) is used widely in marketing for product design, segmentation, and marketing strategy. We propose and test a new “polyhedral” question-design method that adapts each respondent’s choice sets based on previous answers by that respondent. Individual adaptation appears promising because, as demonstrated in the aggregate customization literature, question design can be improved based on prior estimates of the respondent’s partworths – information that is revealed by respondents’ answers to prior questions. The algorithm designs questions that quickly reduce the set of partworths that are consistent with the respondent’s choices. Recent polyhedral “interior-point” algorithms provide the rapid solutions necessary for real-time computation.

To identify domains where individual adaptation is promising (and domains where it is not), we evaluate the performance of polyhedral CBC methods with Monte Carlo experiments. We vary magnitude (response accuracy), respondent heterogeneity, estimation method, and question-design method in a 4×2^3 experiment. The estimation methods are Hierarchical-Bayes estimation (HB) and Analytic-Center estimation (AC). The latter is a new individual-level estimation procedure that is a by-product of polyhedral question design. The benchmarks for individual adaptation are random designs, orthogonal designs, and aggregate customization. The simulations suggest that polyhedral question design does well in many domains, particularly those in which heterogeneity and partworth magnitudes are relatively large. We close by describing an empirical application to the design of executive education programs in which 354 web-based respondents answered stated-choice tasks with four service profiles each. The findings confirm the feasibility of implementing polyhedral CBC methods with actual respondents, test an important design criterion (choice-balance), and provide empirical data on convergence.

1. Introduction

Choice-based conjoint analysis (CBC) describes a class of techniques that are amongst the most widely adopted market research methods. In CBC tasks respondents are presented with two or more product profiles and asked to choose the profile that they prefer (see Figure 1 for an example). This contrasts with other conjoint tasks that ask respondents to provide preference ratings for product attributes or profiles. Because choosing a preferred product profile is often a natural task for respondents consistent with marketplace choice, supporters of CBC have argued that it yields more accurate responses. CBC methods have been shown to perform well when compared against estimates of marketplace demand (Louviere, Hensher, and Swait 2000).

Figure 1

Example of a CBC Task for the Redesign of Polaroid's I-Zone Camera

Question

3

Here is a new set of four cameras. Remember: The features of these options have changed. Click on the white square below to tell us which of the cameras, if any, that you would be willing to purchase.

	Camera M	Camera N	Camera P	Camera Q
Price	\$34.99	\$24.99	\$34.99	
Picture Removal	Manual	Automatic	Automatic	
Picture Taking	2 Step	1 Step	2 Step	
Styling Covers	Changeable	Permanent	Permanent	None of these cameras.
Picture Quality	Option B	Option A	Option B	
Camera Opening	Slide Open	Fixed	Fixed	
Light Selection	3 settings	Feedback	3 settings	
I would purchase: (click only one)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Click on the feature icons for a reminder.

Help
Back
Next

Important academic research investigating CBC methods has sought to improve the design of the product profiles shown to each respondent. This has led to efficiency improvements,

yielding more information from fewer responses. Because an increasing amount of market research is conducted on the Internet, new opportunities for efficiency improvements have arisen. Online processing power makes it feasible to adapt questions based on prior responses. To date the research on adaptive question design has focused on adapting questions based on responses from prior respondents (“aggregate customization”). Efficient designs are customized based on parameters obtained from pretests or from managerial judgment. Examples of aggregate customization methods include Huber and Zwerina (1996), Arora and Huber (2001) and Sandor and Wedel (2001).

In this study we propose a CBC question design method that adapts questions using the previous answers from that respondent (“individual adaptation”). The design of each choice task varies according to that respondent’s selection from prior choice tasks. The approach is motivated in part by the success of aggregate customization, which uses the responses from other respondents to design more efficient questions. The algorithm that we propose focuses on what is not known about partworths (given the respondent’s answers to prior questions) and seeks to reduce quickly the set of partworths that are consistent with the respondent’s choices. To achieve this goal we focus on four design criteria: non-dominance, feasibility, choice balance, and symmetry. In later discussion we also describe an analogy between the proposed algorithm and D-efficiency.

After data are collected with these adaptive questions, partworths can be estimated with standard methods (aggregate random utility or Hierarchical Bayes methods). As an alternative, we propose and test an individual-level estimation method that relies on the analytic center of a feasible set of parameters.

Our proposal differs in both format and philosophy from the other individual-level adaptive conjoint analysis (ACA) methods. We focus on stated-choice data rather than ACA’s metric paired-comparisons and we focus on analogies to efficient design rather than ACA’s utility balance subject to orthogonality goals. Polyhedral methods are also feasible for metric-paired-comparison data (Toubia, Simester, Hauser and Dahan 2003). However, as will become apparent, there are important differences between the metric-paired-comparison algorithm and the algorithm proposed in this paper.

The remainder of the paper is organized as follows. We begin by reviewing existing CBC question design and estimation methods. We next propose a polyhedral approach to the

design of CBC questions. We then evaluate the proposed polyhedral methods using a series of Monte Carlo simulations, where we hope to demonstrate the domains in which the proposed method shows promise (and where existing methods remain best). We compare performance against three question design benchmarks, including an aggregate customization method that uses prior data from either managerial judgment or pretest respondents. Because we expect that individual-adaptation is most promising when responses are accurate and/or respondents are heterogeneous, we compare the four question design methods across a range of customer heterogeneity and response error domains, while also varying the estimation method. We then describe an empirical application of the proposed method to the design of executive education programs at a major university. The paper concludes with a review of the findings, limitations, and opportunities for future research.

2. Existing CBC Question Design and Estimation Methods

To date, most applications of CBC assume that each respondent answers the same set of questions or that the questions are either blocked across sets of respondents or chosen randomly. For these conditions, McFadden (1974) shows that the inverse of the covariance matrix, Σ , of the MLE estimates is proportional to:

$$(1) \quad \Sigma^{-1} = R \sum_{i=1}^q \sum_{j=1}^{J_i} (\bar{z}_{ij} - \sum_{k=1}^{J_i} \bar{z}_{ik} P_{ik})' P_{ij} (\bar{z}_{ij} - \sum_{k=1}^{J_i} \bar{z}_{ik} P_{ik})$$

where R is the effective number of replicates; J_i is the number of profiles in choice set i ; q is the number of choice sets; \bar{z}_{ij} is a binary vector describing the j^{th} profile in the i^{th} choice set; and P_{ij} is the probability that the respondent chooses profile j from the i^{th} choice set. Without loss of generality, we use binary vectors in the theoretical development to simplify notation and exposition. Multi-level features are used in both the simulations and the application.

We can increase precision by decreasing a measure (norm) of the covariance matrix, that is, by either increasing the number of replicates or increasing the terms in the summations of Equation 1. Equation 1 also demonstrates that the covariance of logit-based estimates depends on the choice probabilities, which, in turn, depend upon the partworths. In general, the experimental design that provides the most precise estimates will depend upon the parameters.

Many researchers have addressed choice set design. One common measure is D-efficiency, which seeks to reduce the geometric mean of the eigenvalues of Σ (Kuhfield, Tobias,

and Garratt 1994).¹ If \bar{u} represents the vector of the partworths, then the confidence region for maximum likelihood estimates (\hat{u}) is an ellipsoid defined by $(\bar{u} - \hat{u})' \Sigma^{-1} (\bar{u} - \hat{u})$, Greene (1993, p. 190). The length of the axes of this ellipsoid are given by the eigenvalues of the covariance matrix, so that minimizing the geometric mean of these eigenvalues shrinks the confidence region around the estimates.

Because efficiency depends on the partworths, it is common to assume *a priori* that the stated choices are equally likely. For this paper we will label such designs as “orthogonal” efficient designs. Arora and Huber (2001), Huber and Zwerina (1996), and Sandor and Wedel (2001) demonstrate that we may improve efficiency by using data from either pretests or prior managerial judgment. These researchers improve D-efficiency by “relabeling,” which permutes the levels of features across choice sets, “swapping,” which switches two feature levels among profiles in a choice set, and “cycling,” which is a combination of rotating levels of a feature and swapping them. The procedures stop when no further improvement is possible.² Simulations suggest that these procedures improve efficiency and, hence, reduce the number of respondents that are necessary. Following the literature, we label these designs as “aggregate customization.”

Estimation

In classical logit analysis partworths are estimated with maximum likelihood techniques. Because it is rare that a respondent will be asked to make enough choices to estimate partworth values for each respondent, the data usually are merged across respondents to estimate population-level (or segment-level) partworth values. Yet managers often want estimates for each respondent. Hierarchical Bayes (HB) methods provide (posterior) estimates of partworths for individual respondents by using population-level distributions of partworths to inform individual-level estimates (Allenby and Rossi 1999; Arora, Allenby and Ginter 1998; Johnson 1999; Lenk, et. al. 1996). In particular, HB methods use data from the full sample to iteratively estimate both the posterior means (and distribution) of individual-level partworths and the posterior distribution of those partworths at the population-level. The HB method is based on Gibbs sampling and the Metropolis Hastings Algorithm.

Liechty, Ramaswamy, and Cohen (2001) demonstrate the effectiveness of HB for choice menus, while Arora and Huber (2001) show that it is possible to improve the efficiency of HB estimates with choice-sets designed using Huber-Zwerina relabeling and swapping. In other re-

cent research, Andrews, Ainslie, and Currim (2002, p. 479) present evidence that HB models and finite mixture models estimated from simulated scanner-panel data “recover household-level parameter estimates and predict holdout choice about equally well except when the number of purchases per household is small.”

3. Polyhedral Question Design Methods

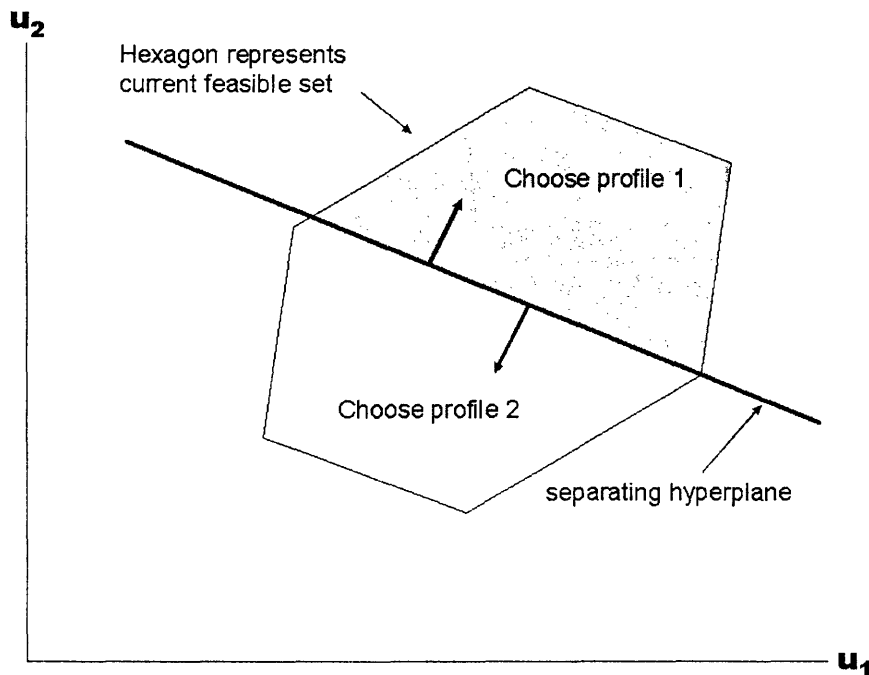
We extend the philosophy of customization by developing algorithms to adapt questions for each respondent. Stated choices by each respondent provide information about parameter values for that respondent that can be used to select the next question(s). In high dimensions (high p) this is a difficult dynamic optimization problem. We address this problem by making use of extremely fast algorithms based on projections within the interior of polyhedra (much of this work started with Karmarkar 1984). In particular, we draw on the properties of bounding ellipsoids discovered in theorems by Sonnevend (1985a, 1985b) and applied by Freund (1993), Nesterov and Nemirovskii (1994), and Vaidja (1989).

We begin by illustrating the intuitive ideas in a two-dimensional space with two product profiles ($J_i = 2$) and then generalize to a larger number of dimensions and multichotomous choice (the simulations and the application are based on multichotomous choice.) The axes of the space represent the partworths (utilities) associated with two different product attributes, u_1 and u_2 . A point in this space has a value on each axis, and will be represented by a vector of these two partworths. The ultimate goal is to estimate the point in this space (or distribution of points) that best represents each respondent. The question-design goal is to focus precision toward the points that best represent each respondent. This goal is not unlike D-efficiency, which seeks to minimize the confidence region for estimated partworths. Without loss of generality we scale all partworths in the figures to be non-negative and bounded from above. Following convention, the partworth associated with the least-preferred level is set arbitrarily to zero.³

Suppose that we have already asked $i-1$ stated-choice questions and suppose that the hexagon (polyhedron) in Figure 2 represents the partworth vectors that are consistent with the respondent’s answers. Suppose further that the i^{th} question asks the respondent to choose between two profiles with feature levels \bar{z}_{i1} and \bar{z}_{i2} . If there were no response errors, then the respondent would select Profile 1 whenever $(z_{i11} - z_{i21}) u_1 + (z_{i12} - z_{i22}) u_2 \geq 0$; where z_{ijf} refers to the f^{th} feature of z_{ij} and u_f denotes the partworth associated with the f^{th} feature. This inequality con-

straint defines a separating line or, in higher dimensions, a separating hyperplane. In the absence of response errors, if the respondent's true partworth vector is above the separating hyperplane the respondent chooses Profile 1; if it is below the respondent chooses Profile 2. Thus, the respondent's choice of profiles updates our knowledge of which partworth vectors are consistent with the respondent's preferences, shrinking the feasible polyhedron.

Figure 2
Stated Choice Responses Divide the Feasible Region



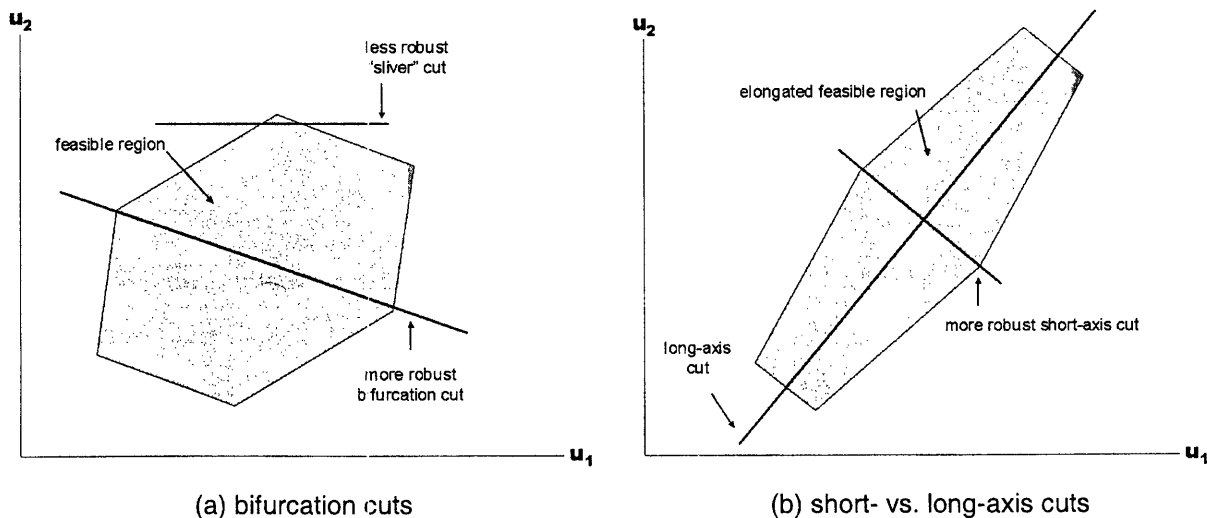
Selecting Questions

Questions are more informative if they reduce the feasible region more rapidly. To implement this goal we adopt four criteria. First, neither profile in the choice set should dominate the other profile. Otherwise, we gain no information when the dominating profile is chosen and the partworth space is not reduced. Second, the separating hyperplane should intersect with and divide the feasible region derived from the first $i-1$ questions. Otherwise, there could be an answer to the i^{th} question that does not reduce the feasible region. A corollary of these criteria is that, for each profile, there must be a point in the feasible region for which that profile is the preferred profile.

The i^{th} question is more informative if, given the first $i-1$ questions, the respondent is equally likely to select each of the J_i profiles. This implies that, a priori, the answer to the i^{th} question should be approximately equally likely – a criterion we call “choice balance.” Choice balance will shrink the feasible region as rapidly as feasible. For example, if the points in the feasible region are equally likely (based on $i-1$ questions) then the predicted likelihood, π_{ij} , of choosing the j^{th} region is proportional to the size of the region. The expected size of the region after the i^{th} question is then proportional to $\sum_{j=1}^2 \pi_{ij}^2$, which is minimized when $\pi_{ij} = \frac{1}{2}$.⁴ Arora and Huber (2001, p. 275) offer a further motivation for choice balance based on D-efficiency. For two product profiles, the inverse covariance matrix, Σ^{-1} , is proportional to a weighted sum of $\pi_{ij}(1 - \pi_{ij})$, which is also maximized for $\pi_{ij} = \frac{1}{2}$.

The choice-balance criterion will hold approximately if we favor separating hyperplanes that go through the center of the feasible polyhedron and cut the feasible region approximately in half. This is illustrated in Figure 3a, where we favor bifurcation cuts relative to “sliver” cuts that yield unequal sized regions. If the separating hyperplane is a bifurcation cut, both the non-domination and feasibility criteria are satisfied automatically.

Figure 3
Comparing Cuts and the Resulting Feasible regions



However, not all bifurcation cuts are equally robust. Suppose that the current feasible region is elongated, as in Figure 3b, and we must decide between many separating hyperplanes,

two of which are illustrated. One cuts along the long axis and yields long thin feasible regions while the other cuts along the short axis and yields feasible regions that are more symmetric. The long-axis cut focuses precision where we already have high precision, while the short-axis cut focuses precision where we now have less precision. For this reason, we prefer short-axis cuts to make the post-choice feasible regions reasonably symmetric. We can also motivate this criterion relative to D-efficiency. D-efficiency minimizes the geometric mean of the axes of the confidence ellipsoid – a criterion that tends to make the confidence ellipsoids more symmetric.

For two profiles we favor non-dominance, feasibility, choice balance, and post-choice symmetry if we select profiles such that (a) the separating hyperplanes go through the center of the feasible region and (b) the separating hyperplanes are perpendicular to the longest “axis” of the feasible polyhedron as defined by the first $i-1$ stated choices. To implement these criteria we propose the following heuristic algorithm.

Step 1 Find the center and the longest axis of the polyhedron based on $i-1$ questions.

Step 2 Find the two intersections between the longest axis and the boundary of the polyhedron.

Each intersection point is defined by a partworth vector and the difference between these vectors defines a hyperplane that is perpendicular to the longest axis. Because respondents choose from profiles rather than partworth vectors, we need a third step to identify profiles that yield this separating hyperplane:

Step 3 Select a profile corresponding to each intersection point such that the separating hyperplane divides the region into two approximately equal sub-regions.

The basic intuition remains the same when we extend this heuristic algorithm to more than two profiles ($J_i > 2$), although the geometry becomes more difficult to visualize (some hyperplanes become oblique, but the regions remain equi-probable). We first address this generalization and then describe several implementation issues, including how we use utility maximization to select the profiles in Step 3.

Selecting More than Two Profiles

In a choice task with more than two profiles the respondent's choice defines more than one separating hyperplane. The hyperplanes that define the i^{th} feasible region depend upon the profile chosen by the respondent. For example, consider a choice task with four product profiles, labeled 1, 2, 3, and 4. If the respondent selects Profile 1, then we learn that the respondent prefers Profile 1 to Profile 2, Profile 1 to Profile 3, and Profile 1 to Profile 4. This defines three separating hyperplanes – the resulting polyhedron of feasible partworths is the intersection of the associated regions and the prior feasible polyhedron. In general, J_i profiles yield $J_i(J_i-1)/2$ possible hyperplanes. For each of the J_i choices available to the respondent, J_i-1 hyperplanes contribute to the definition of the new polyhedron. The full set of hyperplanes, and their association with stated choices, define a set of J_i convex regions, one associated with each answer to the stated-choice question.

We extend Steps 1 to 3 as follows. Rather than finding the longest axis, we find the $(J_i/2)$ longest axes and identify the J_i points where the $(J_i/2)$ longest axes intersect the polyhedron. If J_i is odd, we select randomly among the vectors intersecting the $(J_i/2)^{\text{th}}$ longest axis. We associate profiles with each of the J_i partworth vectors by solving the respondent's maximization problem for each vector (as described next). Our solution to the maximization problem assures the hyperplanes go (approximately) through the center of the polyhedron. The approximation arises because we design the profiles from a discrete attribute space. They would pass exactly through the analytic center if the attribute space were continuous.

It is easy to show that such hyperplanes divide the feasible region into J_i collectively exhaustive and mutually exclusive convex sub-regions of approximately equal size (except for the regions' "indifference" borders, which have zero measure). Non-dominance and feasibility remain satisfied and the resulting regions tend toward symmetry. Because the separating hyperplanes are defined by the profiles associated with the partworth vectors (Step 3) not the partworth vectors themselves (Step 2), some of the hyperplanes do not line up with the axes. For $J_i > 2$ the stated properties remain approximately satisfied based on "wedges" formed by the J_i-1 hyperplanes. Later in the paper we examine the effectiveness of the proposed heuristic for $J_i = 4$. Simulations examine overall accuracy and an empirical test examines whether feasibility and choice balance are achieved for real respondents.

Implementation

Implementing this heuristic raises challenges. Although it is easy to visualize (and implement) the heuristic with two profiles in two dimensions, practical CBC problems require implementation with J_i profiles in large p -dimensional spaces with p -dimensional polyhedra and $(p-1)$ -dimensional hyperplane cuts. Furthermore, the algorithm should run sufficiently fast so that there is little noticeable delay between questions.

The first challenge is finding the center of the current polyhedron and the $J_i/2$ longest axes (Step 1). If we define the longest “axis” of a polyhedron as the longest line segment in the polyhedron, then we would need to enumerate all vertices of the polyhedron and compute the distances between the vertices. Unfortunately, for large p this problem is computationally intractable (Gritzmann and Klee 1993); solving it would lead to lengthy delays between questions for each respondent. Furthermore, this definition of the longest axes of a polyhedron may not capture the intuitive concepts that we used to motivate the algorithm.

Instead, we turn to Sonnevend’s theorems (1985a, 1985b) which state that the shape of polyhedra can be approximated with bounding ellipsoids centered at the “analytic center” of the polyhedron. The analytic center is the point that maximizes the geometric mean of the distances to the boundaries. Freund (1993) provides efficient algorithms to find the analytic centers of polyhedra. Once the analytic center is found, Sonnevend’s results provide analytic expressions for the ellipsoids. The axes of ellipsoids are well-defined and capture the intuitive concepts in the algorithm. The longest axes are found with straightforward eigenvalue computations for which there are many efficient algorithms. With well-defined axes it is simple to find the part-worth vectors on the boundaries of the feasible set that intersect the axes (Step 2). We provide technical details in an Appendix.

To implement Step 3 we must define the respondent’s utility maximization problem. We do so in an analogy to economic theory. For each of the J_i utility vectors on the boundary of the polyhedron we obtain the j^{th} profile, \bar{z}_{ij} , by solving:

$$\text{(OPT1)} \quad \max \bar{z}_{ij} \bar{u}_{ij} \quad \text{subject to:} \quad \bar{z}_{ij} \bar{c} \leq M, \text{ elements of } \bar{z}_{ij} \in \{0,1\}$$

where \bar{u}_{ij} is the utility vector chosen in Step 2, \bar{c} are “costs” of the features, and M is a “budget constraint.” We implement (approximate) choice balance by setting \bar{c} equal to the ana-

lytic center of the feasible polyhedron (\bar{u}_{i-1}) computed after the first $i-1$ questions. At optimality the constraint in OPT1 will be approximately binding, which implies that $\bar{z}_{ij}\bar{u}_{i-1} \cong \bar{z}_{ik}\bar{u}_{i-1} \cong M$ for all $k \neq j$. It may be approximate due to the integrality constraints in OPT1 (elements of $\bar{z}_{ij} \in \{0,1\}$). This assures that the $J_i(J_i-1)/2$ separating hyperplanes go (approximately) through the analytic center. The binding constraints are all bifurcations which tends to make the regions approximately equal in size. Finally, we also know that the solution to OPT1 ensures that each of the separating hyperplanes pass through the feasible polyhedron because each profile is preferred at the utility vector to which it corresponds.

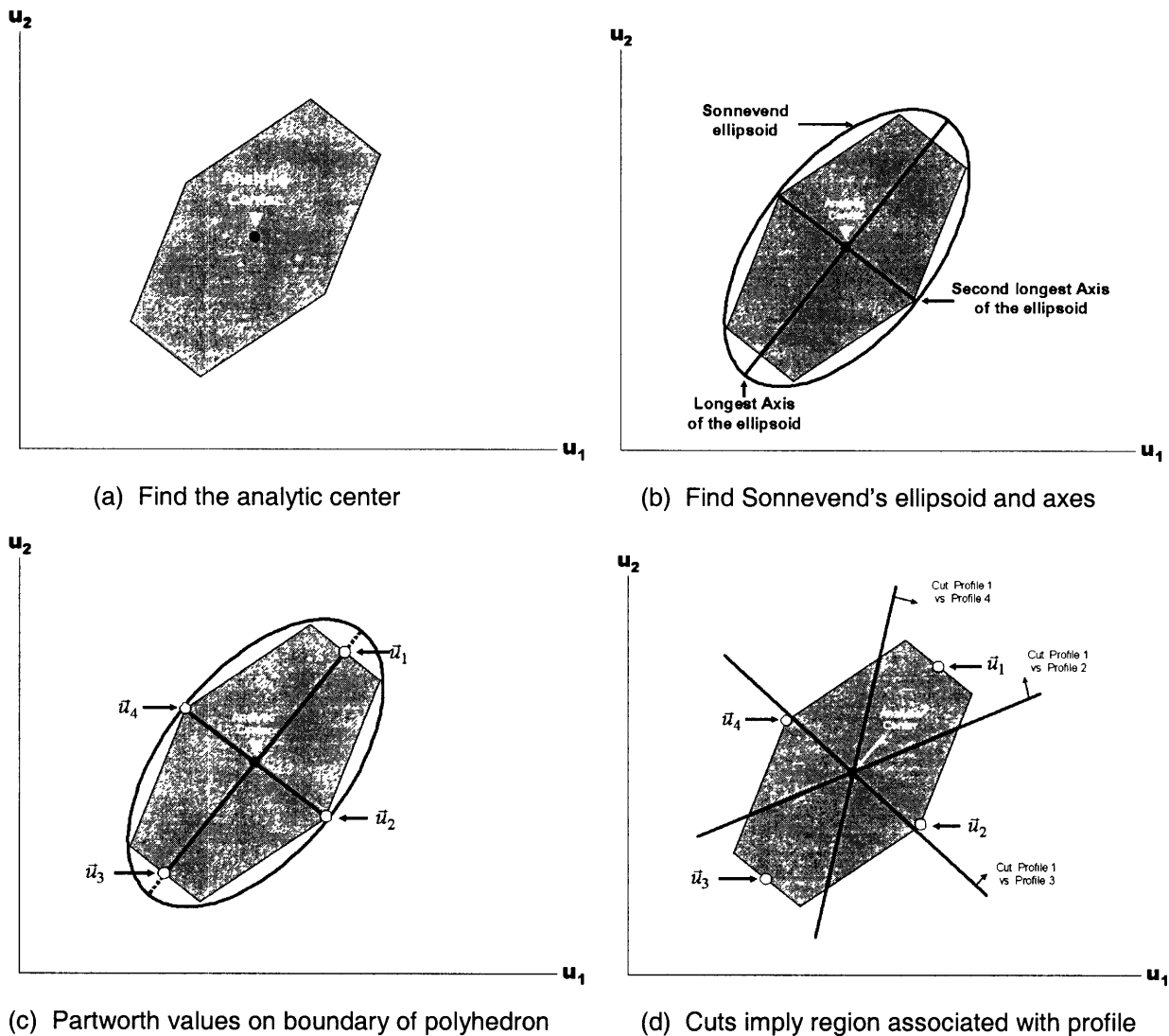
Solving OPT1 for profile selection (Step 3) is a knapsack problem which is well-studied and for which efficient algorithms exist. M is an arbitrary constant that we draw randomly from a compact set (up to m times) until all profiles in a stated-choice task are distinct. If the profiles are not distinct we use those that are distinct. If none of the profiles are distinct then we ask no further questions (in practice this is a rare occurrence in both simulation and empirical situations).

OPT1 also illustrates the relationship between choice balance and utility balance – a criterion in aggregate customization. In our algorithm, the J_i profiles are chosen to be equally likely based on data from the first $i-1$ questions. In addition, for the partworths at the analytic center of the feasible region the utilities of all profiles are approximately equal. However, utility balance only holds at the analytic center, not throughout the feasible region. Thus, while, a priori, the profiles are equally likely to be chosen, it will be rare that the respondent is indifferent among the profiles. Hence, choice balance is unlikely to lead to respondent fatigue and we observed none in our empirical application.

We illustrate the algorithm for $J_i = 4$ with the two-dimensional example in Figure 4. We begin with the current polyhedron of feasible partworth vectors (in Figure 4a). We then use Freund's algorithm to find the analytic center of the polyhedron as illustrated by the black dot in Figure 4a. We next use Sonnevend's formulae to find the equation of the approximating ellipsoid and obtain the $J_i/2$ longest axes (Figure 4b), which correspond to the $J_i/2$ smallest eigenvalues of the matrix that defines the ellipsoid. We then identify J_i target partworth vectors by finding the intersections of the $J_i/2$ axes with the boundaries of the current polyhedron (Figure 4c). Finally, for each target utility vector we solve OPT1 to identify J_i product profiles. The J_i product profiles each imply J_i-1 hyperplanes (illustrated for Profile 1 in Figure 4d). If the respondent

chooses Profile 1, this implies a new smaller polyhedron defined by the separating hyperplanes. As drawn in Figure 4d, one of the hyperplanes is redundant; this is less likely in higher dimensions. Were we to draw all $J_i(J_i-1)/2 = 6$ hyperplanes, they would divide the polyhedron into mutually exclusive and collectively exhaustive convex regions of approximately equal size. We continue for q questions or until OPT1 no longer yields distinct profiles.

Figure 4
Bounding Ellipsoids and the Analytic Center of the Polyhedra



Incorporating Managerial Constraints and other Prior Information

Previous research suggests that prior constraints enhance estimation (Johnson 1999; Srinivasan and Shocker 1973). For example, self-explicated data might constrain the rank order of partworth values across features. Such constraints are easy to incorporate and shrink the feasible polyhedron. Most conjoint analysis studies use multi-level features, some of which are ordinal scaled (e.g., picture quality). For example, if u_{fm} and u_{fh} are the medium and high levels of feature f , we add the constraint, $u_{fm} \leq u_{fh}$, to the feasible polyhedron.⁵ We similarly incorporate information from managerial priors or pretest studies.

Response Errors

In real questionnaires there are likely response errors in stated choices. When there are response errors, the separating hyperplanes are approximations rather than deterministic cuts. For this and other reasons, we distinguish question selection and estimation. The algorithm we propose is a question-selection algorithm. After the data are collected we can estimate the respondents' partworths with most established methods, which address response error formally. For example, polyhedral questions can be used with classical random-utility models or HB estimation. It remains an empirical question as to whether or not response errors counteract the potential gains in question selection due to individual-level adaptation. Although we hypothesize that the criteria of choice balance and symmetry lead to robust stated-choice questions, we also hypothesize that individual-level adaptation will work better when response errors are smaller. We examine these issues in the next section.

Analytic-Center (AC) Estimation

The analytic center of the i^{th} feasible polyhedron provides a natural summary of the information in the first i stated-choice responses. This summary measure is a good working estimate of the respondent's partworth vector. It is a natural byproduct of the question-selection algorithm and is available as soon as each respondent completes the i^{th} stated-choice question. Such estimates might also be used as starting values in HB estimation, as estimates in classical Bayes updating, and as priors for aggregate customization.

Analytic-Center estimates also give us a means to test the ability of the polyhedral algorithm to implement the feasibility and choice-balance criteria. Specifically, if we use the i^{th} AC

estimate to forecast choices for $q > i$ choice sets, it should predict 100% of the first i choices (feasibility) and $(1/J_i)$ percent of the last $q - i$ choices (choice balance). When J_i does not vary with i , the internal predictive percentage should approximately equal $[i + (1/J_i)(q-i)]/q$. We examine this statistic in the empirical application later in the paper.

We can give the AC estimate a statistical interpretation if we assume that the probability of a feasible point is proportional to its distance to the boundary of the feasible polyhedron. In this case, the analytic center maximizes the likelihood of the point (geometric mean of the distances to the boundary).

AC estimates each respondent's partworth vectors based on data only from that respondent. This advantage is also a disadvantage because, unlike Hierarchical Bayes estimation, AC estimation does not use information from other respondents. This suggests an opportunity to improve the accuracy of AC estimates by using data from the population distribution of partworths. While the full development of such an AC algorithm is beyond the scope of the paper, we can test its potential by using the (known) population distribution as a Bayesian prior to update AC estimates. We hypothesize that AC estimates will be less accurate (relative to HB) when the respondents are homogeneous, but that this disadvantage can be offset with the development of an AC Bayesian hybrid.

Incorporating Null Profiles

Many researchers prefer to include a null profile as an additional profile in the choice set (as in Figure 1). Polyhedral concepts generalize readily to include null profiles. If the null profile is selected from choice set i then we add the following constraints: $\bar{z}_{ij}\bar{u} \leq \bar{z}_k^*\bar{u} \quad \forall j, k \neq i$ where \bar{z}_k^* denotes the profile chosen from choice set k (given that the null profile was not chosen in choice set k). Intuitively, these constraints recognize that if the null profile is selected in one choice set, then all of the alternatives in that choice set have a lower utility than the profiles selected in other choice sets (excluding other choice sets where the null was chosen). Alternatively, we can expand the parameter set to include the partworth of an "outside option" and write the appropriate constraints. After incorporating these constraints the question design heuristic (and analytic-center estimation) proceed as described above. We leave practical implementation, Monte Carlo testing, and empirical applications with null profiles to future research.

Metric-Paired-Comparison Questions

Polyhedral methods are also feasible for metric-paired comparison questions. In particular, Toubia et al. (2003) propose a polyhedral method for metric paired-comparison data that also uses Sonnevend's ellipsoids. However, metric-pair and the CBC formats present fundamentally different challenges resulting in very different polyhedral algorithms. For example, each metric-pair question defines equality constraints which reduce the dimensionality of the feasible polyhedron. In the CBC algorithm the inequality constraints do not reduce the dimensionality of the feasible polyhedron. Furthermore, the metric-pairs polyhedron becomes empty after sufficient questions. The metric-pairs algorithm must revert to an alternative question selection method and the metric-pairs analytic-center algorithm must address infeasibility. In the CBC algorithm, the polyhedron always remains feasible. Moreover, because the metric-pairs algorithm identifies the partial profiles directly, the utility-maximization knapsack algorithm is new to the CBC algorithm.

Summary

Polyhedral (ellipsoid) algorithms provide a means to adapt stated-choice questions for each respondent based on that respondent's answers to the first $i-1$ questions. The algorithms are based on the intuitive criteria of non-dominance, feasibility, choice balance, and symmetry and represent an individual-level analogy to D-efficiency. Specifically, the polyhedral algorithm focuses questions on what is not known about the partworth vectors and does so by seeking a small feasible region. Choice balance, symmetry, and the shrinking ellipsoid regions provide analogies to D-efficiency which seeks questions to minimize the confidence ellipsoid for maximum-likelihood estimates.

While both polyhedral question design and aggregate customization are compatible with most estimation methods, including AC estimation, the two methods represent a key tradeoff. Polyhedral question design adapts questions for each respondent but may be sensitive to response errors. Aggregate customization uses the same design for all respondents, but is based on prior statistical estimates that take response errors into account. This leads us to hypothesize that polyhedral methods will have their greatest advantages relative to existing methods (question design and/or estimation) when responses are more accurate and/or when respondents' partworths are more heterogeneous. We next examine individual-level adaptation and AC estima-

tion with Monte Carlo experiments.

4. Monte Carlo Experiments

We use Monte Carlo experiments to investigate whether polyhedral methods show sufficient promise to justify further development and to identify the empirical domains in which the potential is greatest. Monte Carlo experiments are widely used to evaluate conjoint analysis methods, including studies of interactions, robustness, continuity, attribute correlation, segmentation, new estimation methods, and new data-collection methods. In particular, they have proven particularly useful in the first tests of aggregate customization and in establishing domains in which aggregate customization is preferred to orthogonal designs. Monte Carlo experiments offer several advantages for an initial test of new methods. First, with any heuristic, we need to establish computational feasibility. Second, Monte Carlo experiments enable us to explore many domains and control the parameters that define those domains. Third, other researchers can readily replicate and extend Monte Carlo experiments, facilitating further exploration and development. Finally, Monte Carlo experiments enable us to control the “true” partworth values, which are unobserved in studies with actual consumers.

However, Monte Carlo experiments are but the first step in a stream of research. Assumptions must be made about characteristics that are not varied, and these assumptions represent limitations. In this paper we explore domains that vary in terms of respondent heterogeneity, response accuracy (magnitude), estimation method, and question-design method. This establishes a 4×2^3 experimental design. We encourage subsequent researchers to vary other characteristics of the experiments.

Structure of the Simulations

For consistency with prior simulations, we adopt the basic simulation structure of Arora and Huber (2001). Arora and Huber varied response accuracy, heterogeneity, and question-design method in a 2^3 experiment using Hierarchical Bayes (HB) estimation. Huber and Zwerina (1996) had earlier used the same structure to vary response accuracy and question-design with classical estimation, while more recently Sandor and Wedel (2001) used a similar structure to compare the impact of prior beliefs.

The Huber-Zwerina, and Arora-Huber algorithms were aggregate customization methods based on relabeling and swapping. These algorithms work best for stated-choice problems in

which relabeling and swapping are well-defined. We expand the Arora and Huber design to include four levels of four features for four profiles, which ensures that complete aggregate customization and orthogonal designs are possible. Sandor and Wedel included cycling, although they note that cycling is less important in designs where the number of profiles equals the number of feature levels.⁶

Within a feature Arora and Huber choose partworths symmetrically with expected magnitudes of $-\bar{\beta}$, 0, and $+\bar{\beta}$. They vary response accuracy by varying $\bar{\beta}$. Larger $\bar{\beta}$'s imply higher response accuracy because the variance of the Gumbel distribution, which defines the logit model, is inversely proportional to the squared magnitude of the partworths (Ben-Akiva and Lerman 1985, p. 105, property 3). For four levels we retain the symmetric design with magnitudes of $-\bar{\beta}$, $-\frac{1}{3}\bar{\beta}$, $\frac{1}{3}\bar{\beta}$, and $\bar{\beta}$. Arora and Huber model heterogeneity by allowing partworths to vary among respondents according to normal distributions with variance, σ_{β}^2 . They specify a coefficient of heterogeneity as the ratio of the variance to the mean. Specifically, they manipulate low response accuracy with $\bar{\beta} = 0.5$ and high response accuracy with $\bar{\beta} = 1.5$. They manipulate high heterogeneity with $\sigma_{\beta}^2 / \bar{\beta} = 2.0$ and low heterogeneity with $\sigma_{\beta}^2 / \bar{\beta} = 0.5$. Given these values, they draw each respondent's partworths from a normal distribution with a diagonal covariance matrix. Each respondent then answers the stated-choice questions with probabilities determined by a logit model based on that respondent's partworths. Arora and Huber compare question selection using root mean squared error (RMSE). For comparability, we adopt the same criterion and, in addition, report three more intuitive metrics.

We select magnitudes and heterogeneity that represent the range of average partworths and heterogeneity that we might find empirically. While we could find no meta-analyses for these values, we did have available to us data from a proprietary CBC application (D-efficient design, HB estimation) in the software market. The study included data from almost 1,200 home consumers and over 600 business customers. In both data sets $\bar{\beta}$ ranged from approximately -3.0 to $+2.4$. We chose our high magnitude (3.0) from this study recognizing that other studies might have even higher magnitudes. For example, Louviere, Hensher and Swait (2000) report stated-choice estimates (logit analysis) that are in the range of 3.0 and higher.

After selecting $\bar{\beta}$ for high magnitudes, we set the low magnitude $\bar{\beta}$ to the level chosen by Arora and Huber. In the empirical data, the estimated variances ranged from 0.1 to 6.9 and the

heterogeneity coefficient varied from 0.3 to 3.6.⁷ To approximate this range and to provide symmetry with the magnitude coefficient, we manipulated high heterogeneity with a coefficient of three times the mean. Following Arora and Huber we manipulated low heterogeneity as half the mean. We feel that these values are representative of those that might be obtained in practice. Recall that, as a first test of polyhedral methods, we seek to identify domains that can occur in practice and for which polyhedral methods show promise. More importantly, these levels illustrate the directional differences among methods and, hence, provide insight for further development.

Experimental Design

In addition to manipulating magnitude (two levels) and heterogeneity (two levels) we manipulate estimation method (two levels), and question-design method (four levels). The estimation methods are Hierarchical Bayes and Analytic Center estimation. The former is well-established and incorporates information from other respondents in each individual estimate. The latter is the only feasible method, of which we are aware, that provides individual-level estimates using only information from that respondent. The question-design methods are random, orthogonal designs with equally-likely priors, aggregate customization (Arora-Huber), and polyhedral methods. To simulate aggregate customization, we assume that the pretest data are obtained costlessly and, based on this data, we apply the Arora-Huber algorithm. Specifically, we simulate an orthogonal “pretest” that uses the same number of respondents as in the actual study. For the orthogonal design we adopt the Arora-Huber methods as detailed in Huber and Zwerina (1996, p. 310-312).

We set $q = 16$ so that orthogonal designs, relabeling, and swapping are well-defined. Exploratory simulations suggest that the estimates become more accurate as we increase the number of questions, but the relative comparisons of question design and estimation for $q = 8$ and for $q = 24$ provide similar qualitative insights.⁸ For each combination of question design method, estimation method, heterogeneity level and magnitude level we simulated 1,000 respondents.⁹

Practical Implementation Issues

In order to implement the polyhedral algorithm we made two implementation decisions: (1) We randomly drew M up to thirty times ($m = 30$) for the simulations. We believe that the accuracy of the method is relatively insensitive to this decision. (2) Because, prior to the first ques-

tion, the polyhedron is symmetric, we selected the first question by randomly choosing from among the axes.

Other decisions may yield greater (or lesser) accuracy, hence the performance of the polyhedral methods tested in this paper should be considered a lower bound on what is possible with further improvement. For example, future research might use aggregate customization to select the first question. All polyhedral optimization, question selection, and estimation algorithms are described in the Appendix and implemented in Matlab code. The web-based application described later in this paper uses Perl and Html for web-page presentation. All code (and the orthogonal design) are available at mitsloan.mit.edu/vc and is open-source.

Comparative Results of the Monte Carlo Experiments

Table 1 reports four metrics describing the simulation results in a table format similar to Arora and Huber: root mean square error (RMSE); the percentage of respondents for whom each question design method has the lowest RMSE; the “hit rate,” and the average correlation between the true and estimated partworths. The hit rate measures the percentage of times each method predicts the most-preferred profile and is based on 1,000 sets of holdout profiles. We use *italic bold* text to indicate the best question design method for each estimation method within an experimental domain (and any other methods that are not statistically different from the best method). Tables 2a and 2b summarize the best question design method and the best estimation method, respectively, for each metric in each domain. The entries in Tables 2 correspond to the best overall method (question design x estimation) for each domain.

For comparability between estimation methods we first normalized the partworths to a constant scale. Specifically, for each respondent, we normalize both the “true” partworths and the estimated partworths so that their absolute values sum to the number of parameters and their values sum to zero for each feature. In this manner, the RMSEs can be interpreted as a percent of the mean partworths. Within an estimation method, subject to statistical confidence, this scaling does not change the relative comparisons among question design methods. This scaling has the added advantage of making the results roughly comparable in units for the different manipulations of magnitude (response accuracy) and heterogeneity. This scaling addresses two issues. First, Analytic-Center estimation is unique to a positive linear transformation and thus focuses on the relative values of the partworths – as required by many managerial applications. Second, un-

scaled logit analyses confound the magnitude of the stochasticity of observed choice behavior with the magnitude of the partworths. Our scaling enables us to focus on the relative partworths. For volumetric forecasts, we recommend the methods proposed and validated by Louviere, Hensher, and Swait (2000). These methods are well-documented and reliable and have been proven appropriate for matching disparate scales.

Table 1
Monte Carlo Simulation Results

Magnitude and Heterogeneity			RMSE		Percent Best		Hit Rates		Correlations	
Mag	Het	Question Design	HB	AC	HB	AC	HB	AC	HB	AC
Low	High	random	0.892[†]	1.116	26.9[†]	17.2	0.538	0.508	0.656[†]	0.506
		orthogonal	0.904	0.950	21.7	22.1	0.540	0.526	0.651[†]	0.569
		customized	0.883[†]	0.993	27.6[†]	26.9	0.548[†]	0.539[†]	0.660[†]	0.553
		polyhedral	0.928	0.880[†]	23.8[†]	33.8[†]	0.527	0.538[†]	0.632	0.609[†]
Low	Low	random	1.027	1.206	21.9	16.5	0.421	0.403	0.589	0.446
		orthogonal	0.964[†]	1.012[†]	29.6[†]	28.9	0.438[†]	0.425	0.629[†]	0.535[†]
		customized	<i>1.018</i>	1.086	28.7[†]	37.9[†]	0.423	0.436[†]	0.589	0.523[†]
		polyhedral	1.033	1.103	19.6	16.7	0.418	0.403	0.587	0.509
High	High	random	0.595	0.812	23.5	15.0	0.627[†]	0.584	0.813	0.666
		orthogonal	0.815	0.871	4.3	5.8	0.590	0.581	0.715	0.646
		customized	0.611	0.891	31.9	10.0	0.630[†]	0.581	0.802	0.626
		polyhedral	0.570[†]	0.542[†]	40.3[†]	69.2[†]	0.626[†]	0.636[†]	0.819[†]	0.760[†]
High	Low	random	0.446	0.856	31.8	16.3	0.750	0.632	0.903	0.676
		orthogonal	0.692	0.824	2.8	11.7	0.668	0.626	0.804	0.680
		customized	0.769	1.012	18.8	10.1	0.612	0.558	0.698	0.570
		polyhedral	0.418[†]	0.571[†]	46.6[†]	61.9[†]	0.761[†]	0.704[†]	0.912[†]	0.798[†]

[†] Best or not significantly different than best at $p < 0.05$. Note that lower values of RMSE reflect increased accuracy, while higher values on percent best, hit rates, and correlations denote increased accuracy.

Table 2a
Comparison Summary for Question Design

Magnitude	Heterogeneity	RMSE	% Best	Hit Rates	Correlations
Low	High	Random Customized Polyhedral	Polyhedral	Customized	Random Orthogonal Customized
Low	Low	Orthogonal	Orthogonal Customized	Orthogonal Customized	Orthogonal
High	High	Polyhedral	Polyhedral	Polyhedral	Polyhedral
High	Low	Polyhedral	Polyhedral	Polyhedral	Polyhedral

Table 2b
Comparison Summary for Estimation

Magnitude	Heterogeneity	RMSE	% Best	Hit Rates	Correlations
Low	High	AC HB	AC	HB	HB
Low	Low	HB	AC HB	AC HB	HB
High	High	AC	AC	AC	HB
High	Low	HB	HB	HB	HB

Question Design Methods

As hypothesized, polyhedral question design performs well in the high magnitude (high response accuracy) domains. For these domains, polyhedral question design is best, or tied for best, for all metrics and for both estimation methods. These domains favor individual-level adaptation because the design of subsequent questions is based on more accurate information from

the previous answers. When magnitudes are low, the greater response error works against individual adaptation and polyhedral question design does not do as well.

The impact of heterogeneity is more complex. We expect individual-level adaptation to perform well when respondents are heterogeneous. When both magnitudes and heterogeneity are high, conditions favor polyhedral question design and it performs well. In this domain, the accuracy of the data enables individual-level adaptation to be efficient. However, for low magnitudes the disadvantages of high response errors appear to offset the need for customization (high heterogeneity). In this domain, polyhedral question design is best only for AC estimation, perhaps because the two polyhedral methods are complementary – polyhedral question design shrinks the feasible region rapidly making AC estimation more accurate. We expect low heterogeneity to reduce the need for customization and low response accuracy (low magnitude) to work against deterministic customization. In this domain, as predicted, polyhedral methods do not do as well as orthogonal and aggregate customized designs.

Perhaps one surprise in Table 1 is the strong performance of random designs relative to orthogonal designs – especially for HB estimation and when either magnitudes or heterogeneity are high. We believe that this relative performance can be explained by two phenomena. First, orthogonal designs are optimal only when the partworths are zero. For higher magnitudes orthogonal designs are further from optimal (Huber and Zwerina 1996; Arora and Huber 2001). For example, when we compute D-errors for orthogonal and random designs in the high magnitude domains, the D-errors are higher for orthogonal questions than for random questions. Second, HB uses inter-respondent information effectively. As Sandor and Wedel (2003) illustrate, multiple designs are more likely to contribute incremental information about the population. The accuracy of random designs relative to fixed, orthogonal designs is consistent with simulations of differentiated designs as proposed by Sandor and Wedel (2003).

Finally, we note three aspects of Table 1. First, Table 1 replicates the Arora-Huber simulations when the domains and metric are matched – Arora and Huber use HB and report RMSE. In the Arora-Huber simulations, aggregate customization is superior to orthogonal designs when magnitudes are high and when heterogeneity is high. Second, there are ties in Table 1, especially for the low magnitude and high heterogeneity domain, perhaps because high heterogeneity favors customization while low magnitudes make customization more sensitive to errors. Even when we increase the sample size to 1,500 (for RMSE) we are unable to break the ties. In this

domain, for most practical problems, question design appears less critical. Third, performance varies slightly by metric, especially when there are ties and especially for low magnitudes.

Estimation Methods

The findings in Table 1 also facilitate the comparison of estimation methods. We have already noted that AC performs well when matched with polyhedral questions for high heterogeneity. In most other domains (and for most metrics), HB is more accurate. In theory, the advantages of HB come from a number of properties, one of which is the use of population-level data to moderate the individual-level estimates (shrinkage). We expect this to be particularly beneficial when the population is homogenous, because the population-level data provide more information about individual preferences in these domains. This is consistent with Table 1, and may help to explain why AC's relative performance improves when the population is more heterogeneous.

Before we reject AC estimation for the domains in which HB is now superior, we investigate the theoretical potential to improve AC by replacing the AC estimate with a convex combination of the AC estimate and the population mean (shrinkage). If we knew the population mean ($\bar{\beta}$), its variance (σ_{β}^2), and the accuracy of AC (RMSE), then classical Bayes updating provides a formula by which to implement shrinkage. To test shrinkage, we use the known population mean, its variance, and the RMSE from Table 1. With these values we compute a single parameter for each domain, α , with which to weigh the population mean. Using these theoretical α 's, a convex combination of AC and the population mean provides the best overall estimate in all four domains – RMSEs of 0.863, 0.871, 0.510, and 0.403, respectively, the last three are significantly best at the 0.05 level. In practice, we do not know α , so we must estimate it from the data. Optimal estimation is beyond the scope of this paper, but surrogates, such as using either HB or classical logit analysis to estimate α , should perform well.

We conclude that AC shows sufficient promise to justify further development.

Summary of Monte Carlo Experiments

We summarize the results of the Monte Carlo experiments as follows:

- Polyhedral question design shows the most promise when magnitudes are high (response errors are low).

- If magnitudes are low (response errors are high) it may be best not to customize designs; fixed orthogonal questions appear to be more accurate than polyhedral or customized methods.
- HB estimation does well in all domains.
- AC estimation performs well when matched with polyhedral question design and when heterogeneity is high.
- AC estimation shows sufficient promise to justify further development. Preliminary analysis suggests that a modified Bayesian AC estimate is particularly promising if an optimal means can be found to estimate a single parameter, α .

Like many new technologies, we hope polyhedral question design and analytic-center estimation will improve further with use, experimentation, and evolution (Christensen 1998).

5. Application to the Design of Executive Education Programs

Polyhedral methods for CBC have been implemented in at least one empirical application. We describe this application briefly as evidence that it is feasible to implement the proposed methods using actual respondents. Furthermore, the data provide empirical estimates of magnitude and heterogeneity, enable us to test the feasibility and choice-balance criteria, and provide insight on the convergence of AC and HB estimates. Because the managerial environment required that we implement a single question-design method, we cannot compare question-design methods in this first “proof-of-concept” application.

Feature Design and Sample Selection for the Polyhedral Choice-Based Conjoint Study

The application supported the design of new executive education programs for a business school at a major university. The school was the leader in twelve-month executive advanced-degree programs with a fifty-year track record that has produced alumni who include CEOs, prime ministers, and a secretary general of the United Nations. However, the demand for twelve-month programs was shrinking because it has become increasingly difficult for executives to be away from their companies for twelve months. The senior leadership of the school was considering a radical redesign of their programs. As part of this effort, they sought input from potential students. Based on two qualitative studies and detailed internal discussions, the

senior leadership of the school identified eight program features that were to be tested with joint analysis. The features included program focus (3 levels), format (4 levels), class composition (3 levels of interest focus, 4 age categories, 3 types of geographic focus), sponsorship strategy (3 levels), company focus (3 levels), and tuition (3 levels). This $4^2 \times 3^6$ is relatively large for CBC applications (e.g., Huber 1997, Orme 1999), but proved feasible with professional web design. We provide an example screenshot in Figure 5 (with the university logo and tuition levels redacted). Prior to answering these stated-choice questions, respondents reviewed detailed descriptions of the levels of each feature and could access these descriptions at any time by clicking the feature's logo.

Figure 5

Example Web-Based Stated-Choice Task for Executive-Education Study

EP EXECUTIVE PROGRAMS

Please choose

Please examine the following four programs, each described by their features and tuition. Of these four programs, which do you prefer? Click on the circle below the program you would **MOST** prefer. Click the 'Next' button to continue to the next question.

FEATURES	PROGRAM A	PROGRAM B	PROGRAM C	PROGRAM D
Program Focus	Tech-Driven Enterprise	Global Enterprise	Innovative Enterprise	Tech-Driven Enterprise
Program Format	Full-Time Residential	Flexible	Weekend	On-line
Classmates' Background	General Management	Tech. Management	50 - 50 mix	General Management
Classmates' Age	30 - 35 years	35 - 40 years	30-40 years	35 - 45 years
Classmates' Geographic Comp.	75% North American	75% International	50 - 50 mix	75% North American
Classmates' Org. Sponsorship.	Company Sponsored	Self Sponsored	50 - 50 mix	Company Sponsored
Classmates' Company Size	Small Companies	Large Companies	Mix of large and small	Small Companies
Program Tuition				

NEXT >>

After wording and layout was refined in pretests, potential respondents were obtained from the Graduate Management Admissions Council through their Graduate Management Admissions Search Service. Potential respondents were selected based on their age, geographic location, educational goals, and GMAT scores. Random samples were chosen from three strata – those within driving distance of the university, those within a short airplane flight, and those within a moderate airplane flight. Respondents were invited to participate via e-mails from the

Director of Executive Education. As an incentive, respondents were entered in a lottery in which they had a 1-in-10 chance of receiving a university-logo gift worth approximately \$100.

Pretests confirmed that the respondents could comfortably answer twelve stated-choice questions (recall that the respondents were experienced executives receiving minimal response incentives). Of those respondents who began the CBC section of the survey, 95% completed the section. The overall response rate was within ranges expected for both proprietary and academic web-based studies (Couper 2000; De Angelis 2001; Dillman et. al., Sheehan 2001).¹⁰ No significant differences were found between the partworths estimated for early responders and those estimated for later responders. The committee judged the results intuitive, but enlightening, and adequate for managerial decision-making. Based on the results of the conjoint study and internal discussions, the school has redesigned and retargeted its twelve-month programs with a focus on both global leadership and innovation/entrepreneurship. Beginning with the class of 2004, it is adding a new, flexible format to its traditional offerings.

Technical Results

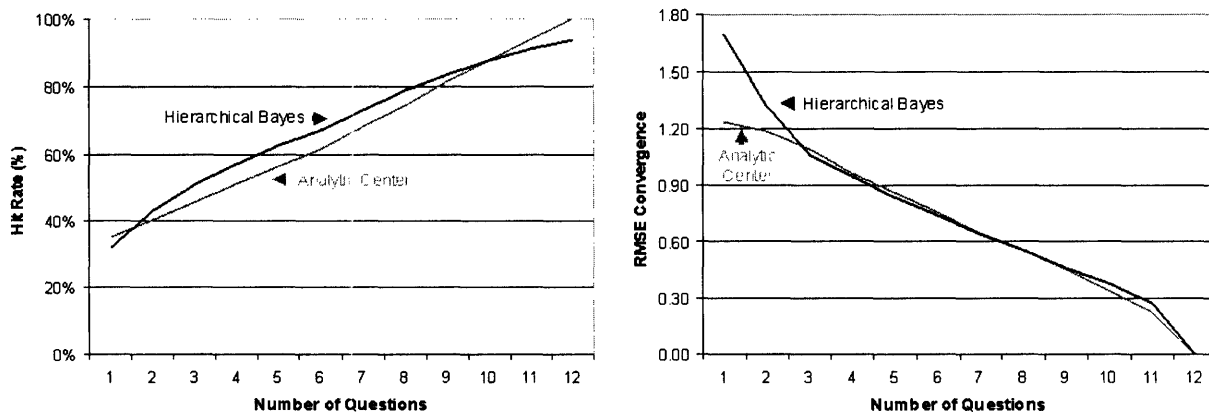
The data provide an opportunity to examine several technical issues. First, we estimate the magnitude and heterogeneity parameters using HB. We obtained estimates of magnitude ($\bar{\beta}$) that ranged from 1.0 to 3.4, averaging 1.5. The heterogeneity coefficient ranged from 0.9 to 3.2, averaging 1.9. These observed magnitude and the observed heterogeneity coefficients span the ranges addressed in the Monte Carlo simulations.

Hit rates are more complex. By design, polyhedral questions select the choice sets that provide maximum information about the feasible set of partworths. If the Analytic-Center (AC) estimates remain feasible through the twelfth question they obtain an internal hit rate of 100% (by design). They remained feasible and this internal hit rate was achieved. This hit rate is not guaranteed for HB estimates, which, nonetheless, do quite well with internal hit rates of 94%. As described earlier in this paper, we can examine internal consistency by comparing the hit rates based on AC estimates from the first i questions to their theoretical value, $[i + \frac{1}{4}(12 - i)]/12$. The fit is almost perfect (adj. $R^2 = 0.9973$), suggesting that polyhedral question design was able to achieve excellent choice balance for the respondents in this study. Because these internal hit rates are not guaranteed for HB, we plot the hit rates for both estimation methods in Figure 6a. On this metric, HB is more concave than AC doing better for low i , but less well for high i .

To gain further insight we adopt an evaluation method used by Kamakura and Wedel (1995, p. 316) to examine how rapidly estimates converge to their final values. Following their structure, we compute the convergence rates as a function of the number of stated-choice tasks (i) using scaled RMSE to maintain consistency with Arora and Huber (2001) and with the Monte Carlo simulations. From question 3 onward, AC and HB are quite close achieving roughly equal convergence. HB does less well for $i = 1$ and 2, most likely because individual-level variation (in a population with moderately high heterogeneity) counterbalances the benefit to HB of the population-level data.

Figure 6

Empirical Hit Rates and RMSE Convergence for Executive-Education Study



(a) Hit Rates

(b) RMSE Convergence

Based on this initial application we conclude that adaptive polyhedral choice-based questions are practical and achieve both feasibility and choice-balance. Analytic-Center estimation appears to be comparable to HB and is deserving of further study, perhaps with the Bayesian hybrids suggested earlier in the text.

6. Conclusions and Research Opportunities

Research on stated-choice question design suggests that careful selection of choice sets has the potential to increase accuracy and reduce costs by requiring fewer respondents, fewer questions, or both. This is particularly true in choice-based conjoint analysis because the most efficient design depends upon the true partworth values. In this paper we explore whether the

success of aggregate customization can be extended to individual-level adaptive question design. We propose heuristics for designing profiles for each choice set. We then rely on new developments in dynamic optimization to implement these heuristics. As a first test, we seek to identify whether or not the proposed methods show promise in at least some domains. It appears that such domains exist. Like many proposed methods, we do not expect polyhedral methods to dominate in all domains and, indeed, they do not. However, we hope that by identifying promising domains we can inspire other researchers to explore hybrid methods and/or improve the heuristics.

While polyhedral methods are feasible empirically and show promise, many challenges remain. For example, we might allow fuzzy constraints for the polyhedra. Such constraints might provide greater robustness at the expense of precision. Future simulations might explore other domains including non-diagonal covariance structures, probit-based random-utility models, mixtures of distributions, and finite mixture models. Recently, Ter Hofstede, Kim, and Wedel (2002) demonstrated that self-explicated data could improve HB estimation for full-profile conjoint analysis. Polyhedral estimation handles such data readily – hybrids might be explored that incorporate both self-explicated and stated-choice data. Future developments in dynamic optimization might enable polyhedral algorithms that look many steps ahead.

We close by recognizing research on other optimization algorithms for conjoint analysis. Evgeniou, Boussios, and Zacharia (2003) propose “support vector machines (SVMs)” to balance complexity of interactions with fit. They are currently exploring hybrids based on SVMs and polyhedral methods.

Endnotes

1. This is equivalent to maximizing the p^{th} root of the determinant of Σ^{-1} . Other norms include A-efficiency, which maximizes the trace of Σ^{-1}/p , and G-efficiency, which maximizes the maximum diagonal element of Σ^{-1} .

2. The Huber-Zwerina and Arora-Huber algorithms maximize $\det \Sigma^{-1}$ based on the mean partworths and do so by assuming the mean is known from pretest data (or managerial judgment). Sandor and Wedel (2001) include, as well, a prior covariance matrix in their calculations. They then maximize the expectation of $\det \Sigma^{-1}$, where the expectation is over the prior subjective beliefs.

3. In the application described later in the paper we use warm-up questions to identify the lowest level of each feature (a common solution to this issue).

4. For J_i profiles, equally-sized regions also maximize entropy, defined as $-\sum_j \pi_{ij} \log \pi_{ij}$. Formally, maximum entropy is equal to the total information obtainable in a probabilistic model (Hauser 1978, Theorem 1, p. 411).

5. In the theoretical derivation we used binary features without loss of generality for notational simplicity. An ordinal multi-level feature constraint is mathematically equivalent to a constraint linking two binary features.

6. In the designs that we use, the efficiency of the Sandor and Wedel algorithm is approximately equal to the efficiency of the Huber and Zwerina algorithm.

7. There was also an outlier with a mean of 0.021 and a variance of 0.188 implying a heterogeneity coefficient of 9.0. Such cases are clearly possible, but less likely to represent typical empirical situations. Researchers who prefer a unitless metric for heterogeneity can rescale using the standard deviation rather than the variance. For our two-level manipulation, directional differences are the same.

8. The estimates at $q = 16$ are approximately 25% more accurate than those at $q = 8$ and the estimates at $q = 24$ are approximately 12% more accurate than those at $q = 16$.

9. The simulations are based on ten sets of one hundred respondents. To reduce unnecessary variance, the same true partworths for each of the 1,000 respondents are used for each question design method.

10. Couper (2000, p. 384) estimates a 10% response rate for open-invitation studies and a 20-25% response rate for studies with pre-recruited respondents. De Angelis (2001) reports “click-through” rates of 3-8% from an e-mail list of 8 million records. In a study designed to optimize response rates, Dillman, et al. (2001) compare mail, telephone, and web-based surveys. They obtain a response rate of 13% for the web-based survey. Sheehan (2001) reviews all published studies referenced in Academic Search Elite, Expanded Academic Index, ArticleFirst, Lexis-Nexis, Psychlit, Sociological Abstracts, ABI-Inform, and ERIC and finds response rates dropping at 4% per year. Sheehan’s data suggest an average response rate for 2002 of 15.5%. The response rate in the Executive Education study was 16%.

References

- Allenby, Greg M. and Peter E. Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, (March/April), 57-78.
- Andrews, Rick L., Andrew Ainslie, and Imran S. Currim (2002), "An Empirical Comparison of Logit Choice Models with Discrete versus Continuous Representations of Heterogeneity," *Journal of Marketing Research*, 39, (November), 479-487.
- Arora, Neeraj, Greg M. Allenby and James L. Ginter (1998), "A Hierarchical Bayes Model of Primary and Secondary Demand," *Marketing Science*, 17, 1, 29-44.
- and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28, (September), 273-283.
- Ben-Akiva, Moshe and Steven R. Lerman (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, (Cambridge, MA: MIT Press).
- Christensen, Clayton (1998). *The Innovator's Dilemma : When New Technologies Cause Great Firms to Fail* (Boston, MA: Harvard Business School Press).
- Couper, Mick P. (2000), "Web Surveys: Review of Issues and Approaches," *Public Opinion Quarterly*, 64, 464-494.
- De Angelis, Chris (2001), "Sampling for Web-based Surveys," 22nd Annual Marketing Research Conference, Atlanta, GA, (September 23-26).
- Dillman, Don A., Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, and Jodi Berck (2001), "Response Rate and Measurement Differences in Mixed Mode Surveys Using Mail, Telephone, Interactive Voice Response, and the Internet," (Pullman, WA: Social and Economic Sciences Research Center, Washington State University).
- Evgeniou, Theodoros, Constantinos Boussios, and Giorgos Zacharia (2003), "Generalized Robust Conjoint Estimation," Working Paper, (Fontainebleau, France: INSEAD), May.
- Freund, Robert (1993), "Projective Transformations for Interior-Point Algorithms, and a Super-linearly Convergent Algorithm for the W-Center Problem," *Mathematical Programming*, 58, 385-414.

- , Robert, R. Roundy, and M.J. Todd (1985), "Identifying the Set of Always-Active Constraints in a System of Linear Inequalities by a Single Linear Program," WP 1674-85, Sloan School of Management, MIT.
- Greene, William H. (1993), *Econometric Analysis, 2E*, (Englewood Cliffs, NJ: Prentice-Hall, Inc.)
- Gritzmann P. and V. Klee (1993), "Computational Complexity of Inner and Outer J-Radii of Polytopes in Finite-Dimensional Normed Spaces," *Mathematical Programming*, 59(2), pp. 163-213.
- Hadley, G. (1961), *Linear Algebra*, (Reading, MA: Addison-Wesley Publishing Company, Inc.)
- Hauser, John R. (1978), "Testing the Accuracy, Usefulness and Significance of Probabilistic Models: An Information Theoretic Approach," *Operations Research*, Vol. 26, No. 3, (May-June), 406-421.
- Huber, Joel (1997), "What we Have Learned from 20 Years of Conjoint Research," Research Paper Series, (Sequim, WA: Sawtooth Software, Inc.).
- and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33, (August), 307-317.
- Johnson, Richard (1999), "The Joys and Sorrows of Implementing HB Methods for Conjoint Analysis," Research Paper Series, (Sequim, WA: Sawtooth Software, Inc.).
- Kamakura, Wagner and Michel Wedel (1995), "Life-Style Segmentation with Tailored Interviewing," *Journal of Marketing Research*, 32, (August), 308-317.
- Karmarkar, N. (1984), "A New Polynomial Time Algorithm for Linear Programming," *Combinatorica*, 4, 373-395.
- Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, 31, 4, (November), 545-557.
- 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, 173-191.
- Liechty, John, Venkatram Ramaswamy, Steven Cohen (2001), "Choice-Menus for Mass Customization: An Experimental Approach for Analyzing Customer Demand With an Application to a Web-based Information Service," *Journal of Marketing Research*, 38, 2, (May).

- Louviere, Jordan J., David A. Hensher, and Joffre D. Swait (2000), *Stated Choice Methods: Analysis and Application*, (New York, NY: Cambridge University Press).
- McFadden, Daniel (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, P. Zarembka, ed., (New York: Academic Press), 105-142.
- Nesterov, Y. and A. Nemirovskii (1994), "Interior-Point Polynomial Algorithms in Convex Programming," SIAM, Philadelphia.
- Orme, Bryan (1999), "ACA, CBC, or Both?: Effective Strategies for Conjoint Research," Working Paper, Sawtooth Software, Sequim, WA.
- Sandór, Zsolt and Michel Wedel (2001), "Designing Conjoint Choice Experiments Using Managers' Prior Beliefs," *Journal of Marketing Research*, 38, 4, (November), 430-444.
- and ---- (2003), "Differentiated Bayesian Conjoint Choice Designs," working paper, (Ann Arbor, MI: University of Michigan Business School).
- Sheehan, Kim (2001), "E-mail Survey Response Rates: A Review," *Journal of Computer-Mediated Communications*, 6, 2, (January).
- Sonnevend, G. (1985a), "An 'Analytic' Center for Polyhedrons and New Classes of Global Algorithms for Linear (Smooth, Convex) Programming," *Proceedings of the 12th IFIP Conference on System Modeling and Optimization*, Budapest.
- (1985b), "A New Method for Solving a Set of Linear (Convex) Inequalities and its Applications for Identification and Optimization," Preprint, Department of Numerical Analysis, Institute of Mathematics, Eötvös University, Budapest, 1985.
- Srinivasan, V. and Allan D. Shocker (1973), "Linear Programming Techniques for Multidimensional Analysis of Preferences," *Psychometrika*, 38, 3, (September), 337-369.
- Ter Hofstede, Frenkel, Youngchan Kim, and Michel Wedel (2002), "Bayesian Prediction in Hybrid Conjoint Analysis," *Journal of Marketing Research*, 39, (May), 253-261.
- Toubia, Olivier, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, forthcoming.
- Vaidja, P. (1989), "A Locally Well-Behaved Potential Function and a Simple Newton-Type Method for Finding the Center of a Polytope," in: N. Megiddo, ed., *Progress in Mathematical Programming: Interior Points and Related Methods*, Springer: New York, 79-90.

Appendix: Mathematics of Polyhedral Methods for CBC Analysis

This appendix is designed to be self-contained. Related math programming involved in finding an interior point, the analytic center, and the Sonnevend ellipsoid are presented in detail by Toubia et. al. (2003) in their metric-pairs algorithm. We include the modified math programming formulations here for completeness. We caution readers that there are important differences between the stated-choice formulations as detailed in this paper and the metric-pair formulations presented by Toubia, et. al. The stated-choice algorithm and the knapsack problem obviously do not arise in the metric-pairs setting.

Definitions and Assumptions

It is helpful to begin with several definitions:

- u_f the f^{th} parameter of the respondent's partworth function where $u_f \geq 0$ is the high level of the f^{th} feature (we assume binary features without loss of generality) and $\sum_{f=1}^p u_f = 100$
- p the number of (binary) features
- \vec{u} the $p \times 1$ vector of parameters
- r the number of externally imposed constraints, of which $r' \leq r$ are inequality constraints.
- \vec{z}_{ij} the $1 \times p$ vector describing the j^{th} profile in the i^{th} choice set, where $j=1$ indexes the respondent's choice from each set
- X the $q(J-1) \times p$ matrix of $\vec{x}_{ij} = \vec{z}_{i1} - \vec{z}_{ij}$ for $i = 1$ to q and $j = 2$ to J (to simplify notation, we drop the i subscript from J_i).

Inequality constraints are incorporated by adding slack variables. For example if there are multiple levels and $u_m \leq u_h$, then $u_h = u_m + v_{hm}$ with $v_{hm} \geq 0$. If there were no errors, the respondent's choices would imply $X\vec{u} \geq \vec{0}$ where $\vec{0}$ is a vector of 0's. We add slack variables and augment \vec{u} such that $X\vec{u} = \vec{0}$. We incorporate the additional constraints by augmenting these equations so that \vec{u} and X include r' additional slack variables and r additional equations. This forms a polyhedron, $\mathbf{P}_{\text{CBC}} = \{ \vec{u} \in \mathfrak{R}^{p+q(J-1)+r'} \mid X\vec{u} = \vec{a}, \vec{u} \geq \vec{0} \}$ where \vec{a} contains non-zero elements due to the external constraints. We begin by assuming that \mathbf{P}_{CBC} is non-empty, that X is full-rank, and that no j exists such that $u_j = 0$ for all \vec{u} in \mathbf{P}_{CBC} .

Interior Point Math Program

To find a feasible interior point, solve the following linear program (see Freund, Roundy and Todd 1985):

$$\max_{\vec{u}, \vec{y}, \theta} \sum_{f=1}^{p+q(J-1)+r'} y_f, \quad \text{subject to: } X\vec{u} = \theta \vec{a}, \quad \theta \geq 1, \quad \vec{u} \geq \vec{y} \geq \vec{0}, \quad \vec{y} \leq \vec{e}$$

where \vec{e} is a vector of 1's. Let $(\vec{u}^*, \vec{y}^*, \theta^*)$ denote a solution. If $\vec{y}^* > \vec{0}$, then $\theta^{*-1} \vec{u}^*$ is an interior point of \mathbf{P}_{CBC} . If $y_f^* = 0$, then $u_f = 0$ for all $\vec{u} \in \mathbf{P}_{\text{CBC}}$. If the linear program is infeasible, then \mathbf{P}_{CBC} is empty.

Analytic Center Math Program

Solve the following math program:

$$\max \sum_{f=1}^{p+q(J-1)+r'} \ln(u_f), \quad \text{subject to: } X\bar{u} = \bar{a}, \quad \bar{u} > \bar{0}$$

We do so using an algorithm developed by Freund (1993) that begins with the feasible point \bar{u}^0 that was found earlier. At each iteration we set $\bar{u}^{t+1} = \bar{u}^t + \alpha^t \bar{d}^t$ where \bar{d}^t is found using the following quadratic approximation of the objective function:

$$\sum_{f=1}^{p+q(J-1)+r'} \ln(u_f + d_f^t) \approx \sum_{f=1}^{p+q(J-1)+r'} \ln(u_f) + \sum_{f=1}^{p+q(J-1)+r'} \left(\frac{d_f^t}{u_f} - \frac{d_f^t{}^2}{2u_f^2} \right)$$

If U^t is a diagonal matrix of the u_f^t 's, then \bar{d}^t solves:

$$\max \bar{e}^T (U^t)^{-1} \bar{d} - (\gamma_2) \bar{d}^T (U^t)^{-2} \bar{d}, \quad \text{subject to: } X\bar{d} = \bar{0}$$

Using the Karush-Kuhn-Tucker (KKT) conditions, $\bar{d}^t = \bar{u}^t - (U^t)^2 X^T [X(U^t)^2 X^T]^{-1} \bar{a}$. If $\| (U^t)^{-1} \bar{d}^t \| < 0.25$, \bar{u}^t is already close to optimal and we set $\alpha^t = 1$. Otherwise we find the optimal α^t with a line search. The program continues to convergence at $\bar{\bar{u}}$.

If \mathbf{P}_{CBC} is empty we employ the error modeling procedure in Toubia, et. al. (2003). Note however that \mathbf{P}_{CBC} will not be empty if CBC questions are chosen with the polyhedral algorithm. If X is not full rank, $X(U^t)^2 X^T$ might not invert. There are two practical solutions: (1) select questions such that X is full rank or (2) make X full rank by removing redundant rows (see Toubia, et. al. 2003). If when searching for feasibility we identify some f 's for which $u_f = 0$ for all $\bar{u} \in \mathbf{P}_{\text{CBC}}$, we can find the analytic center of the remaining polyhedron by removing those f 's and setting $u_f = 0$ for those indices.

The Longest Axes of the Sonnevend Ellipsoid

If $\bar{\bar{u}}$ is the analytic center and $\bar{\bar{U}}$ is the corresponding diagonal matrix, then Sonnevend (1985a, 1985b) demonstrates that $\mathbf{E} \subseteq \mathbf{P}_{\text{CBC}} \subseteq \mathbf{E}_{p+q(J-1)+r'}$ where, $\mathbf{E} = \{ \bar{u} \mid X\bar{u} = \bar{a}, \sqrt{(\bar{u} - \bar{\bar{u}})^T \bar{\bar{U}}^{-2} (\bar{u} - \bar{\bar{u}})} \leq 1 \}$ and $\mathbf{E}_{p+q(J-1)+r'}$ is constructed proportional to \mathbf{E} by replacing 1 with $(p+q(J-1)+r')$. Because we are interested only in the direction of the longest axes of the ellipsoids, we can work with the simpler of the proportional ellipsoids, \mathbf{E} . Let $\bar{g} = \bar{u} - \bar{\bar{u}}$, then the longest axis is a solution to:

$$\max \bar{g}^T \bar{g} \quad \text{subject to: } \bar{g}^T \bar{\bar{U}}^{-2} \bar{g} \leq 1, \quad X\bar{g} = \bar{0}$$

Using the KKT conditions, the solution to this problem is the eigenvector of the matrix,

$(\bar{U}^{-2} - X^T (XX^T)^{-1} X\bar{U}^{-2})$, that is associated with its smallest positive eigenvalue. The direction of the next longest axis is given by the eigenvector associated with the second smallest eigenvalue, etc.

Selecting Profiles for Target Partworth Values

We select the values of the \bar{u}_{ij} 's for the next question ($i = q+1$) based on the longest axes. Each axis provides two target values. For odd J we randomly select from target values derived from the $[(J+1)/2]^{\text{th}}$ eigenvector. To find the extreme estimates of the parameters, \bar{u}_{ij} , we solve for the points where $\bar{u}_{i1} = \bar{\bar{u}} + \alpha_1 \bar{g}_1$, $\bar{u}_{i2} = \bar{\bar{u}} - \alpha_2 \bar{g}_2$, $\bar{u}_{i3} = \bar{\bar{u}} + \alpha_3 \bar{g}_3$, and $\bar{u}_{i4} = \bar{\bar{u}} - \alpha_4 \bar{g}_4$ intersect \mathbf{P}_{CBC} (the generalization to $J \neq 4$ is straightforward). For each α we do this by increasing α until the first constraint in \mathbf{P}_{CBC} is violated. To find the profiles in the choice set we select, as researcher determined parameters, feature costs, \bar{c} , and a budget, M . Without such constraints, the best profile is trivially the profile with all features set to their high levels. Subject to this budget constraint, we solve the following knapsack problem with dynamic programming.

$$\text{(OPT1)} \quad \max \bar{z}_{ij} \bar{u}_{ij} \quad \text{subject to:} \quad \bar{z}_{ij} \bar{c} \leq M, \text{ elements of } \bar{z}_{ij} \in \{0,1\}$$

For multi-level features we impose constraints on OPT1 that only one level of each feature is chosen. In the algorithms we have implemented to date, we set $\bar{c} = \bar{\bar{u}}$ and draw M from a uniform distribution on $[0, 50]$, redrawing M (up to thirty times) until all four profiles are distinct. If distinct profiles cannot be identified, then it is likely that \mathbf{P}_{CBC} has shrunk sufficiently for the managerial problem. For null profiles, extend the constraints accordingly, as described in the text.

Chapter 4: Non-Deterministic Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis

Abstract

Polyhedral methods have been introduced recently both for metric paired-comparison conjoint analysis (Toubia, Simester, Hauser and Dahan 2003) and for choice-based conjoint analysis (Toubia, Hauser, Simester 2004). Although these methods appear promising in simulation tests as well as in field experiments, they systematically underestimate response error. In particular, polyhedral questionnaires are typically designed as if responses contained no error.

In this paper I generalize the polyhedral method for choice-based conjoint analysis introduced by Toubia et al. (2004), and allow response error to be taken into account in the design as well as the estimation of the conjoint questionnaire. More precisely, I consider a hypothetical process in which each conjoint question would be randomly discarded with a positive probability. With this process, the posterior distribution on the partworths is a mixture of uniform distributions supported by polyhedra. I generalize the design criteria of choice balance and post-choice symmetry to mixtures of polyhedra. The deterministic polyhedral method of Toubia et al. (2004) becomes a special case in which the mixture contains only one distribution.

I use simulations to test the validity of this non-deterministic approach. The simulations suggest that when response error is high, the non-deterministic approach improves both polyhedral question selection and polyhedral estimation.

1. Introduction

Polyhedral methods have been introduced recently both for metric paired-comparison conjoint analysis (Toubia, Simester, Hauser and Dahan 2003) and for choice-based conjoint analysis (Toubia, Hauser, Simester 2004). These methods appear promising in simulation tests as well as in field experiments. They are based on a new approach to conjoint analysis, relying on the identification of feasible sets of parameters consistent with the respondent's answers.

However one major limitation of the polyhedral methods introduced so far is their treatment of response error. In particular, response error is systematically underestimated, and polyhedral questionnaires are typically designed as if responses contained no error. Consequently, simulation experiments suggest that polyhedral methods are most useful when response error is low, while other methods appear better when response error is high.

In this paper I offer another interpretation of the polyhedral method for choice-based conjoint analysis introduced by Toubia et al. (2004). This new interpretation allows generalizing the method to take response error into account. The "deterministic" method of Toubia et al. (2004) becomes a special case of the broader set of methods introduced in this paper.

The paper is structured as follows. In Section 2, I briefly review the deterministic polyhedral method of Toubia et al. (2004). In Section 3, I extend this method to take response error into account. I test the performance of the new method in Section 4, using simulations. Section 5 concludes and suggests avenues for future research.

2. Basic polyhedral methods for Choice-Based Conjoint Analysis

In this section I briefly review the method proposed by Toubia, Hauser and Simester (2004).

Choice-based Questions as sets of constraints

Toubia et al. (2004) propose viewing the answers to a choice-based conjoint questionnaire as a set of inequality constraints. In particular, if the vectors \mathbf{x}_j and \mathbf{x}_k represent the j^{th} and k^{th} alternatives in a choice question, and if the vector \mathbf{U} represents the respondent's partworths, then the respondent's utility for alternatives j and k are respectively $\mathbf{x}_j \cdot \mathbf{U} + \varepsilon_j$ and $\mathbf{x}_k \cdot \mathbf{U} + \varepsilon_k$ where ε_j and ε_k represent response error. The respondent will choose alternative j over alternative k only if $(\mathbf{x}_j - \mathbf{x}_k) \cdot \mathbf{U} + (\varepsilon_j - \varepsilon_k) \geq 0$. Each choice between J alternatives can be represented by a set of $(J-1)$

such inequality constraints on U and ε . The overall information available on U and ε is represented by the following set of constraints: $X.U + \varepsilon \geq 0$, $U \geq 0$, $e'.U = 100$ where X represents all the binary comparisons implied by the conjoint questionnaire, $U \geq 0$ results from setting to 0 the partworth of the lowest level of each attribute, and e is a vector of 1's ($e'.U = 100$ is a scaling constraint that is imposed without loss of generality).

Polyhedra as feasible regions

Consider the following optimization problem:

Minimize $\|\varepsilon\|$

subject to: $X.U + \varepsilon \geq 0$, $U \geq 0$, $e'.U = 100$

This optimization problem reflects a typical minimum distance estimation procedure. However, this problem usually does not have a unique solution. (Non-uniqueness of the solution is more likely if the number of choice questions is moderately low compared to the number of parameters.) In this case, the optimal ε is equal to 0 and the set of solutions in U is defined by $\Omega = \{U: X.U \geq 0, U \geq 0, e'.U = 100\}$. The method of Toubia et al. (2004) relies on the observation that Ω is a well-known, well studied mathematical object called a polyhedron. As noted earlier, this polyhedron represents all U that are feasible when there is no response error ($\varepsilon = 0$).

Question design

Let Ω_k be the polyhedron corresponding to the feasible region after question k . The next feasible polyhedron Ω_{k+1} will be defined by the intersection of Ω_k and the set of points satisfying the constraints corresponding to the $(k + 1)^{\text{st}}$ question. Let us assume that the $(k + 1)^{\text{st}}$ question consists of a choice between J alternatives. Potential answers to the question divide Ω_k into J collectively exhaustive and mutually exclusive subpolyhedra $\Omega_{k+1}^1, \Omega_{k+1}^2, \dots, \Omega_{k+1}^J$, such that Ω_{k+1} is equal to Ω_{k+1}^j if and only if the respondent chooses alternative j in question $(k + 1)$. Toubia et al. (2004) adopt two criteria when designing choice questions (they mention two additional criteria, which are implied by choice balance):

Choice balance: $\Omega_{k+1}^1, \Omega_{k+1}^2, \dots, \Omega_{k+1}^J$ should be of similar sizes.

Post-choice symmetry: $\Omega_{k+1}^1, \Omega_{k+1}^2, \dots, \Omega_{k+1}^J$ should be as spherical as possible.

The first criterion minimizes the expected size of the next feasible polyhedron Ω_{k+1} , while the second criterion makes uncertainty on U similar in all dimensions. These two criteria are operationalized by the two following principles:

Principle 1: A respondent with a utility vector equal to the center of Ω_k should be indifferent between all J alternatives.

Principle 2: $\Omega_{k+1}^1, \Omega_{k+1}^2, \dots, \Omega_{k+1}^J$ should divide Ω_k along its longest axes.

Toubia et al. (2004) implement these principles by adopting the following procedure, repeated for all questions k :

- a. Compute the analytic center of Ω_k, AC_k . The analytic center is the point that maximizes the geometric mean of the distance to the boundaries of the polyhedron.
- b. Approximate Ω_k by an ellipse E_k centered at AC_k using Sonnevend's (1985a, b) theorems.
- c. Find the $J/2$ longest axes of E_k (if J is odd find the $(J+1)/2$ longest axes).
- d. Define $U^1 \dots U^J$ as the intersections between the longest axes of the ellipse E_k and the polyhedron Ω_k .
- e. Solve the J following knapsack problems ($j = 1 \dots J$):
 Maximize $x \cdot U^j$
 subject to $x \cdot AC_k \leq M$
 where M is a randomly drawn constant.
- f. The solutions to these knapsack problems, $x^{*1} \dots x^{*J}$, are the alternatives presented in the next choice question.

In the above procedure, Principle 1 is approximately satisfied by using $x \cdot AC_k \leq M$ as the constraint in the Knapsack problems in step *e* (at optimality the constraints will be approximately binding, resulting in $x^{*1} \cdot U \sim x^{*2} \cdot U \sim \dots \sim x^{*J} \cdot U \sim M$). Principle 2 is approximately satisfied by choosing $U^1 \dots U^J$ as the intersection between the polyhedron and the longest axes of its approximating ellipse.

Estimation

Choice-based conjoint questionnaires designed by a polyhedral method can be estimated using any estimation procedure. The question-selection procedure provides an alternative estimate after question k – the analytic center of the feasible polyhedron, AC_k . (Analytic Center estimation can be applied to any set of choice questions, and is not restricted to polyhedral questionnaires.)

Performance

Toubia et al. (2004) evaluate the performance of both their question selection algorithm and of Analytic Center (AC) estimation, using the same simulation design as Arora and Huber (2001). This 2x2 design varies response accuracy and respondent heterogeneity. Polyhedral question selection is compared to randomly generated questions, orthogonal designs, and aggregate customization designs (Huber and Zwerina 1996, Arora and Huber 2001). Analytic center estimation is compared to hierarchical Bayes.

The results of the simulations can be summarized as follows (see Tables 1 to 3 in Toubia et al. 2004 for details):

- Polyhedral question selection achieves an improvement over traditional methods when response error is low. However it does not perform better than traditional methods when response error is high.
- Analytic center estimation does not perform as well as hierarchical Bayes when the population is homogeneous, achieves similar performance when heterogeneity is high and response error is high, and achieves superior performance when heterogeneity is high and response error is low.

The fact that the polyhedral methods (question selection as well as estimation) proposed by Toubia et al. (2004) perform relatively better when response error is low might be due to their deterministic nature. In particular, recall that the definition of the polyhedron relies on the assumption that there is no response error. In the next section I attempt to improve those methods by developing a more general framework that allows taking response error into account.

3. Taking response error into account

Polyhedra and probability density functions

Let Ω be a polyhedron. Let us define P_Ω as the uniform probability density function supported by Ω :

$$P_\Omega(x) = 0 \text{ if } x \notin \Omega$$

$$P_\Omega(x) = P_\Omega(y) \text{ for all } (x,y) \in \Omega^2$$

If we assume a flat (uniform) prior on U , if Ω is our feasible polyhedron, and if we assume that there is no response error, it is easy to show that P_Ω is the posterior distribution of U . In particular, the posterior distribution is 0 for all points outside of the polyhedron, and it is proportional to the prior distribution for all points inside the polyhedron.

Hence, the center of the feasible polyhedron can be interpreted as the posterior expected value of U . More importantly, minimizing the size of the polyhedron can be viewed as minimizing the posterior variance of U , and making the feasible polyhedron as spherical as possible can be viewed as making the posterior variance of U similar in all dimensions.

Introduction of response error

One question

Let Ω_0 be the initial polyhedron defined by the identifying constraints $\{U \geq 0, e' \cdot U = 100\}$. For ease of exposition, let us first consider the special case in which we present the respondent with only one choice question between J alternatives. Without loss of generality, let us assume that the respondent chooses alternative 1. Let Ω_1 be the intersection of Ω_0 with the set of points satisfying the constraints corresponding to the selection of the alternative 1. Ω_1 would be the new polyhedron under the deterministic approach of Toubia et al. (2004).

Suppose now that instead of systematically adding the constraints corresponding to the respondent's choice, we were to take them into account only with probability α , and to simply ignore the question with probability $1 - \alpha$. In this case, our next feasible polyhedron would be Ω_1 with probability α , and it would remain Ω_0 with probability $1 - \alpha$. Our posterior distribution on U would then be a mixture of two distributions, $\alpha P_{\Omega_1} + (1 - \alpha) P_{\Omega_0}$.

I later explore how response error can be captured by appropriately choosing the parameter α . Ignoring a question is different from assuming that the respondent made an error on this

question. Let us denote by α' the probability that the respondent's choice actually matches his or her truly preferred alternative. α will be chosen as a function of α' such that ignoring the question with probability $(1 - \alpha)$ will be approximately equivalent (i.e., lead to the same posterior distribution) to assuming that the respondent made an error with probability $(1 - \alpha')$.

Generalization to multiple questions

Let us now generalize the procedure to longer questionnaires. Let us still consider a process that ignores each question with probability $(1 - \alpha)$. After k questions, the distribution of U would be:

$$f(U) = \frac{\sum_s w_s P_{\Omega_s}}{\sum_s w_s} \quad (1)$$

where the summation is over all subsets s of $\{1 \dots k\}$. In particular, a set of constraints can be associated to each subset s of questions, resulting in a polyhedron Ω_s and a corresponding distribution P_{Ω_s} . The weight $w_s = \alpha^{|s|}(1 - \alpha)^{k - |s|}$ is the probability that s would be the subset of questions taken into account in this hypothetical process ($|s|$ is the number of elements in s). We set $w_s = 0$ if the polyhedron Ω_s is empty.³³

The deterministic approach is a special case of Equation 1, in which $\alpha = 1$.

More complex priors

We have assumed so far that the prior was uniform and characterized by the distribution P_{Ω_0} corresponding to the polyhedron $\Omega_0 = \{U: U \geq 0, e' \cdot U \geq 100\}$. However, any mixture of polyhedra could be used to characterize the initial prior. For example, a normal prior on U could be approximated by a mixture of uniform distributions with increasingly large supports and small weights.

Estimation

An estimate of U is provided by:

³³ This is why we have to normalize the weights in Equation 1 and divide them by $\sum_s w_s$. If w_s were not set to 0 when the polyhedron is empty, then Equation 1 would simply be $f(U) = \sum_s w_s P_{\Omega_s}$.

$$\hat{U} = \frac{\sum_s w_s AC_s}{\sum_s w_s} \text{ where } AC_s \text{ is the analytic center of } \Omega_s.$$

Question design

The two basic principles for question selection used by Toubia et al. (2004) can be applied to the non-deterministic framework proposed here. In particular, a respondent with a utility vector equal to the posterior expected value of U after question k should be approximately indifferent between the alternatives in the $(k+1)^{\text{st}}$ choice set. In addition, the mixtures of polyhedra associated with each alternative in this choice set should divide the initial mixture $\sum_s w_s \Omega_s$ along its “longest axes.” This raises the issue of defining and computing the longest axes of a mixture of polyhedra.

Longest axis of a mixture of polyhedra

Let v_s be the longest axis of the polyhedron Ω_s . Intuitively, the longest axis of the mixture $\sum_s w_s \Omega_s$ should be aligned with the longest axes of the most likely polyhedra in the mixture. More precisely, the longest axis should capture, or “summarize,” the directions of the most likely polyhedra in the mixture. I define the longest axis of the mixture of polyhedra as the vector v^* that maximizes $\sum_s w_s (v_s' \cdot v)^2$. (This expression is a norm on the set of inner products $\{v_s' \cdot v\}_s$.)

Let us define V as the $2^k \times p$ matrix (p being the dimension of U) obtained by stacking the transposed longest axes of each polyhedron in the mixture (each row of V is a transposed longest axis). Let us define Π as the $2^k \times 2^k$ diagonal matrix with elements corresponding to the probabilities associated with each subset (i.e., $\Pi_{ss} = w_s$). We have: $\sum_s w_s (v_s' \cdot v)^2 = v'^* V' \Pi V v$. The vector v^* maximizing $\sum_s w_s (v_s' \cdot v)^2$ is then simply the eigenvector associated with the largest eigenvalue of $V' \Pi V$. (This matrix being symmetric, positive semi-definite, its eigenvalues are all real and non-negative.) I simply define the second longest axis as the eigenvector associated with the second largest eigenvalue, and so on and so forth. By construction, the longest axes are

orthogonal to one another, as they are in the case of a single ellipse. (The procedure just defined is analogous to factor analysis.)

Detailed algorithm

Question selection is summarized as follows:

- a. Compute $\hat{U} = \frac{\sum_s w_s AC_s}{\sum_s w_s}$
- b. Approximate each polyhedron in the mixture by an ellipse and compute the corresponding longest axis v_s .
- c. Find the $J/2$ longest axes of the mixture (if J is odd find the $(J+1)/2$ longest axes)
- d. Define $U^1 \dots U^J$ as the intersections between the longest axes of the mixture and the polyhedron $\Omega_0 = \{U: U \geq 0, e' \cdot U \geq 100\}$
- e. Solve the following J knapsack problems ($j = 1 \dots J$):
 Maximize $x \cdot U^j$
 subject to $x \cdot \hat{U} \leq M$
 where M is a randomly drawn constant.³⁴
- f. The solutions to these problems, $x^{*1} \dots x^{*J}$, are the alternatives presented in the next choice question.

Practical Implementation

When k questions have been asked to the respondent, the distribution of U is a mixture of 2^k polyhedra. As k gets large, the computation of the analytic centers and longest axes of each polyhedron in the mixture takes too long to solve between questions in a web-based setting. However aspects of the structure of the problem can be exploited to reduce the amount of computations required between questions.

First, it is unnecessary to compute the analytic centers and longest axes of all 2^k polyhedra in the mixture after question k : if the analytic centers and longest axes of the polyhedra involved after question $(k-1)$ have been saved, the computation of only 2^{k-1} additional polyhedra

³⁴ In the simulations in this paper, M was drawn up to 30 times, until all solutions to the Knapsack problems were different. If all solutions to the Knapsack problems were similar after 30 draws, the questionnaire stopped. If there were only $K < J$ different solutions after 30 draws, these K options were the alternatives presented to the respondent.

is required. Indeed, the set of 2^k polyhedra can be divided into those in which question k is ignored and those in which it is taken into account. The analytic center and longest axis of any polyhedron in which question k is ignored have already been computed before question k was asked, and can hence be reused.

Second, the computation of the analytic center of a polyhedron involves two steps, one of which can almost always be bypassed in our case. The first step consists in finding an interior point of the polyhedron by solving a Linear Program; the second step consists in applying Newton's method using the interior point as a starting point. After question k , the polyhedron $\Omega_{\{1\dots k\}}$ obtained by taking all k questions into account is a subset of all the other polyhedra in the mixture. Hence the analytic center of $\Omega_{\{1\dots k\}}$ is an interior point of all these polyhedra. This can be exploited by first computing the analytic center of $\Omega_{\{1\dots k\}}$ and using it as a starting point in Newton's method for the other polyhedra. This appears to reduce by a factor of approximately two the time required for the computation of the analytic centers.

Despite these simplifying techniques, the computations required between questions quickly becomes too time consuming. To address this issue, I approximate the distribution of U by a mixture of a subset of the 2^k polyhedra. In particular, I sort the probability weights w_s in decreasing order and compute the analytic centers and longest axes of the polyhedra starting with the most likely ones, until a preset computing time limit is reached. In the following simulations, this time limit was 1 second.

More precisely, if S_{limit} is the subset of polyhedra used to approximate the mixture, then

\hat{U} in step a above is approximated with $\frac{\sum_{s \in S_{limit}} w_s AC_s}{\sum_{s \in S_{limit}} w_s}$, and the longest axis of the mixture in step c

is approximated by the vector maximizing $\sum_{s \in S_{limit}} w_s \cdot (v_s' \cdot v)^2$, instead of $\sum_s w_s \cdot (v_s' \cdot v)^2$.

Computation of α

This paper introduces response error by considering a process in which each question would be ignored with probability $(1-\alpha)$. This process was adopted because it is congruent with polyhedral methods. A more direct introduction of response error would result in a mixture of convex and non-convex sets for which no convenient approximation is known.

However, our choice of α should make this process equivalent to one in which there exists response error and we adjust our posterior distribution accordingly. For example, we should be more likely to ignore a question when response error is larger (i.e., α should be smaller). The link between α and response error should also reflect the fact that ignoring a question is different from recognizing that the respondent made an error on this question. An error from the respondent does tell us that the selected alternative is not the one with the highest “true” Utility, which in itself is informative.

We now derive the appropriate α . Let α_k' be the probability that the respondent selected his or her truly preferred alternative in question k . For simplicity, let us consider a particular noise structure in which the respondent chooses his or her “truly” preferred alternative with probability α_k' and chooses each other $J-1$ alternatives with equal probability $\frac{1-\alpha_k'}{J-1}$. Note that in the following simulations, choices were made according to the logistic probabilities, making this assumption an approximation.³⁵ Let $p(U)$ be the mixture of uniform distributions corresponding to our prior distribution on U before question k . Applying Bayes' rule, the posterior distribution of any vector U is proportional to $\alpha_k' p(U)$ if U is consistent with the answer to question k and it is proportional to $\frac{(1-\alpha_k') \cdot p(U)}{J-1}$ if U is inconsistent with that answer. Both posterior densities are proportional to $p(U)$ and the ratio is:

$$(2) \quad \frac{(J-1) \cdot \alpha_k'}{1-\alpha_k'}$$

Let us now consider the process introduced in this paper in which we ignore question k with probability $1-\alpha_k$. The posterior distribution of U is then the mixture $\alpha_k p_2(U) + (1-\alpha_k)p(U)$ where $p_2(U)$ is the posterior distribution obtained if the prior is $p(U)$ and if the constraints corresponding to question k are taken into account. In particular, $p_2(U) = 0$ if U is outside of the corresponding feasible region Ω , and is equal to $p(U) / P(\Omega)$ where $P(\Omega) = \int_{\Omega} p(U) dU$ if U is in Ω . The

posterior of a point U consistent with the answer to question k is then $\frac{\alpha_k \cdot p(U)}{P(\Omega)} + (1-\alpha_k) \cdot p(U)$.

³⁵ In real situations both this assumption and the assumption of logistic probabilities are approximations, and which one better describes actual consumer choices is an empirical question.

The posterior of a point U inconsistent with that answer is $(1 - \alpha_k)p(U)$. Both posterior densities are again proportional to $p(U)$, and the ratio is now equal to:

$$(3) \quad \frac{\alpha_k / P(\Omega) + (1 - \alpha_k)}{1 - \alpha_k}$$

Let us consider the average case in which $P(\Omega) = 1/J$, which is the target probability dictated by the choice balance criterion (Principle 1). With $P(\Omega) = 1/J$ the ratios (2) and (3) are equal if:

$$(4) \quad \frac{J \cdot \alpha_k + (1 - \alpha_k)}{1 - \alpha_k} = \frac{(J - 1) \cdot \alpha'_k}{1 - \alpha'_k} \Leftrightarrow \alpha_k = \frac{J \cdot \alpha'_k - 1}{J - 1}$$

If α_k satisfies Equation 4 and if choice balance is satisfied, then ignoring question k with probability $(1 - \alpha_k)$ would yield the same posterior distribution as realizing that the respondent made an error on question k with probability α'_k .³⁶ In the case where $P(\Omega)$ is different from $1/J$, i.e., choice balance is not achieved, the two processes will be only approximately equivalent.

For simplicity, in the simulations reported in this paper, I chose the same α for all respondents and all questions. In particular, a parameter α' was estimated based on a pretest, and α was computed as a function of α' according to Equation 4.

α' was computed using the following procedure:

- a. An estimate of the population mean of U , \hat{U}_{pop} , was obtained based on a pretest (100 respondents assigned to an orthogonal design in the simulations).
- b. A set of R random questions ($R = 100$ in the simulations) was generated and the logistic probabilities based on \hat{U}_{pop} were computed for each question. (probabilities $p_r^1 \dots p_r^J$ corresponded to the J alternatives of the r^{th} question). The probability that the respondent chose his or her truly preferred alternative on question r was $\alpha'_r = \max \{ p_r^1 \dots p_r^J \}$. An

initial estimate of α' , $\hat{\alpha}'_{ini}$, was obtained as $(\sum_{r=1}^R \alpha'_r) / R$.

³⁶ Equation 4 assumes that $\alpha'_k > 1/J$, i.e., the choice question is informative.

- c. N respondents ($N = 100$ in the simulations) were simulated, all with $U = \hat{U}_{pop}$, using $\alpha = \frac{J \cdot \hat{\alpha}'_{ini} - 1}{J - 1}$ in Equation 1. For each respondent, the logistic probabilities were saved. α' was estimated as in step *b*.

Note that step *b* was necessary because an initial value of α was required in order to design the questions in step *c*. In theory, step *c* should have been repeated until convergence, i.e., until the value of α' used to design the questionnaires was the same as the one obtained from the choice probabilities. However, the value of α' obtained from the choice probabilities did not appear to be very sensitive to the one used to design the questionnaires, and only one iteration was performed in the simulations. Thus, the simulations will understate the performance of an algorithm that is repeated until convergence.

Note also that the pretest information required by this procedure is the same as that required by traditional aggregate customization (Huber and Zwerina 1996, Arora and Huber 2001). This facilitated the comparison between the two methods.

4. Simulations

Design

The design of the simulations reported in this section was similar to that adopted by Arora and Huber (2001) and Toubia et al. (2004). As in Toubia et al. (2004), the number of features was 4, the number of levels per feature was 4, and the number of alternatives per choice set, J , was 4. This allowed designing orthogonal designs using the procedure described by Arora and Huber (2001). As in prior studies, the partworths of the four levels came from a normal distribution with mean $-\beta$, $-\beta/3$, $\beta/3$, and β respectively, and with standard deviation σ_β (the covariance between partworths was 0). The magnitude of the partworths (influencing the level of response error) was controlled by the parameter β , with high magnitude represented by $\beta = 3$, and low magnitude by $\beta = 0.5$. The heterogeneity of the population was controlled by the ratio σ_β^2/β , which was set to 3 in the case of high heterogeneity, and 0.5 in the case of low heterogeneity. For each combination of magnitude and heterogeneity, I simulated 10 sets of 100 respondents.

In addition to manipulating magnitude (two levels) and heterogeneity (two levels), I also varied question design (5 different question selection methods) and estimation (3 different esti-

mation methods). More precisely, for each simulated respondent, I considered five sets of questions:

1. A random design
2. An orthogonal design
3. A design obtained by aggregate customization, using swapping and relabeling (Arora and Huber 2001).
4. A questionnaire designed by the deterministic polyhedral method of Toubia et al. (2004).
5. A questionnaire designed by the non-deterministic polyhedral method proposed in this paper.

For each respondent, each set of questions was estimated by

1. Hierarchical Bayes (using the 100 respondents in the corresponding set).
2. The deterministic AC estimation method of Toubia et al. (2004).
3. The non-deterministic AC estimation method proposed in this paper.

Note that both aggregate customization and the non-deterministic polyhedral method described in this paper require a prior estimate of the population mean. For each of the 10 sets of respondents, this was provided by the population estimate obtained when running hierarchical Bayes on the orthogonal designs.

Results

I compare the methods on their ability to recover respondents' true utility vectors. Table 1 reports the root mean squared error (RMSE) achieved by each combination of question design and estimation method. Tables 2 and 3 summarize the best question-design method and the best estimation method, respectively. For comparability between estimation methods, I normalized the partworths to a constant scale. Specifically, for each respondent, I normalized both the true partworths and the estimated partworths so that their absolute values summed to the number of parameters and their values summed to zero for each feature (Toubia et al. 2004). This scaling addresses two issues. First, polyhedral estimation is unique to a positive linear transformation and thus focuses on the *relative* values of the partworths. Second, unscaled logit analyses confound the magnitude of the stochasticity of observed choice behavior with the magnitude of the partworths.

Question-design methods

Not surprisingly, taking response error into account is more useful to polyhedral methods when response error is higher. In particular, when response error is high (low magnitude) and heterogeneity is high, non-deterministic polyhedral questions coupled with non-deterministic AC estimation outperform deterministic polyhedral questions. However, when response error is large and heterogeneity is low, although non-deterministic AC estimation outperforms deterministic AC estimation, deterministic and non-deterministic polyhedral questions coupled with non-deterministic AC estimation achieve similar performance, and do not perform as well as orthogonal designs coupled with non-deterministic AC estimation.

When response error is low (high magnitude), the performance of non-deterministic polyhedral questions is usually slightly lower than that of deterministic polyhedral questions. However, non-deterministic polyhedral questions still outperform the other question-design methods, with any of the three estimation methods.

I hypothesize that the difference in performance between deterministic and non-deterministic polyhedral questions in the low response error cases is due to the fact only a subset of the polyhedra in the mixture was considered (see “practical implementation”). Future research might explore the sensitivity of performance to this approximation, and allow isolating the benefits of the conceptual framework proposed in this paper from the limitations introduced by implementation choices. Certainly with improved computer speed and memory (Moore’s Law) the need for this approximations will disappear.

Estimation methods

Toubia et al. (2004) note that deterministic AC estimation performs better (relative to hierarchical Bayes) when heterogeneity is high. Their simulations also reveal that it performs better when response error is low (magnitude is high). Table 1 indicates that, when response error is high, deterministic AC estimation can be significantly improved by taking error into account. In particular, non-deterministic AC estimation outperforms both deterministic AC estimation and HB estimation in the low magnitude / high heterogeneity and in the low magnitude / low heterogeneity conditions.

When response error is low (high magnitude), non-deterministic AC estimation does not always perform as well as deterministic AC estimation. As with question-design methods, the difference in performance between deterministic and non-deterministic AC estimation when re-

response error is low is possibly due to the approximations in the implementation of the non-deterministic method.

Table 1
Simulation Results

<i>Magnitude</i>	<i>Heterogeneity</i>	<i>Question Design</i>	<i>HB</i>	<i>RMSE</i>	
				<i>Deterministic AC</i>	<i>Non-deterministic AC</i>
Low	High	Random	0.746	0.878	0.672
		Orthogonal	0.751	0.771	0.620
		Customized	<i>0.696*</i>	0.831	0.635
		Polyhedral – Deterministic	0.801	<i>0.712*</i>	0.646
		Polyhedral – Non deterministic	0.874	0.998	<i>0.604*</i>
Low	Low	Random	0.864	1.151	0.782
		Orthogonal	<i>0.775*</i>	<i>0.898*</i>	<i>0.708*</i>
		Customized	0.796	0.993	0.776
		Polyhedral – Deterministic	0.867	0.941	0.746
		Polyhedral – Non deterministic	0.855	1.224	0.750
High	High	Random	0.540	0.667	0.670
		Orthogonal	0.697	0.824	0.696
		Customized	0.503	0.813	0.760
		Polyhedral – Deterministic	<i>0.478*</i>	<i>0.393*</i>	<i>0.459*</i>
		Polyhedral – Non deterministic	0.489	0.415	<i>0.463*</i>
High	Low	Random	0.380	0.664	0.662
		Orthogonal	0.647	0.813	0.699
		Customized	0.383	1.008	0.893
		Polyhedral – Deterministic	<i>0.328*</i>	<i>0.427*</i>	0.461
		Polyhedral – Non deterministic	<i>0.325*</i>	0.445	<i>0.444*</i>

* Best or not significantly different from best at $p < 0.05$. The comparisons are done across question designs for each estimation method and for each combination of Magnitude / Heterogeneity

Table 2
Comparison Summary for Question Design

<i>Magnitude</i>	<i>Heterogeneity</i>	<i>Lowest RMSE</i>
Low	High	Non-deterministic Polyhedral
Low	Low	Orthogonal
High	High	Deterministic Polyhedral
High	Low	Deterministic Polyhedral / Non-deterministic Polyhedral

Table 3
Comparison Summary for Estimation

<i>Magnitude</i>	<i>Heterogeneity</i>	<i>Lowest RMSE</i>
Low	High	Non-deterministic AC
Low	Low	Non-deterministic AC
High	High	Deterministic AC
High	Low	HB

Summary of the simulation results

The simulations suggest that:

- Taking response error into account improves the performance of polyhedral question selection when response error is high and heterogeneity is high. However the improvement requires coupling non-deterministic polyhedral question design with non-deterministic AC estimation.

Taking response error into account improves the performance of polyhedral estimation when response error is high. In this case, non-deterministic AC estimation also outperforms hierarchical Bayes estimation.

5. Conclusions and Future Research

In their conclusion, Toubia et al. (2004) note: "...many challenges remain. For example, we might allow fuzzy constraints for the polyhedra." In this paper I attempt to address this challenge by generalizing their approach and allowing response error to be taken into account in the design as well as the estimation of polyhedral questionnaires. One difficulty comes from the necessity to deal only with convex polyhedra, which would not be satisfied if response error were introduced directly. Instead, I consider a hypothetical process in which each conjoint question would be randomly discarded with a positive probability. With this process, the posterior distribution on the partworths is a mixture of uniform distributions supported by polyhedra. The deterministic polyhedral method becomes a special case in which the mixture contains only one distribution. I generalize the criteria of choice balance and post-choice symmetry used by Toubia et al. (2004) to mixtures of polyhedra. The probability of discarding a question is chosen such that the hypothetical process under consideration yields a posterior distribution that is approximately equal to the one obtained if response error were directly taken into account.

I use simulations to test the validity of this non-deterministic approach. The simulations suggest that when response error is high, the non-deterministic approach improves both polyhedral question selection and polyhedral estimation.

There are many exciting areas for future research. I suggest two compelling directions. First, in the current implementation, the prior distribution of the partworths (before the beginning of the questionnaire) is uniform. It would be interesting to explore more sophisticated priors, in particular mixtures of uniform distributions that approximate more complex distributions. Second, although the simulations are promising, it is an important future direction to validate the method using a field experiment. This work has begun.

References

- Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28, (September), 273-283.
- Huber, Joel and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33, (August), 307-317.
- Sonnevend, G. (1985a), "An 'Analytic' Center for Polyhedrons and New Classes of Global Algorithms for Linear (Smooth, Convex) Programming," in *Control and Information Sciences*, Vol. 84. Berlin: Springer Verlag, 866-876.
- _____ (1985b), "A New Method for Solving a Set of Linear (Convex) Inequalities and its Applications for Identification and Optimization," reprint, Department of Numerical Analysis, Institute of Mathematics, Eötvös University, Budapest.
- Toubia, Olivier, John R. Hauser, and Duncan I. Simester (2004), "Polyhedral Methods for Adaptive Choice-based Conjoint Analysis," *Journal of Marketing Research*, 41, (February), 116-131.
- _____, Duncan Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, Vol. 22, No. 3 (Summer), 273-303.

Chapter 5: Properties of Preference Questions: Utility Balance, Choice Balance, Configurators, and M-Efficiency

Abstract

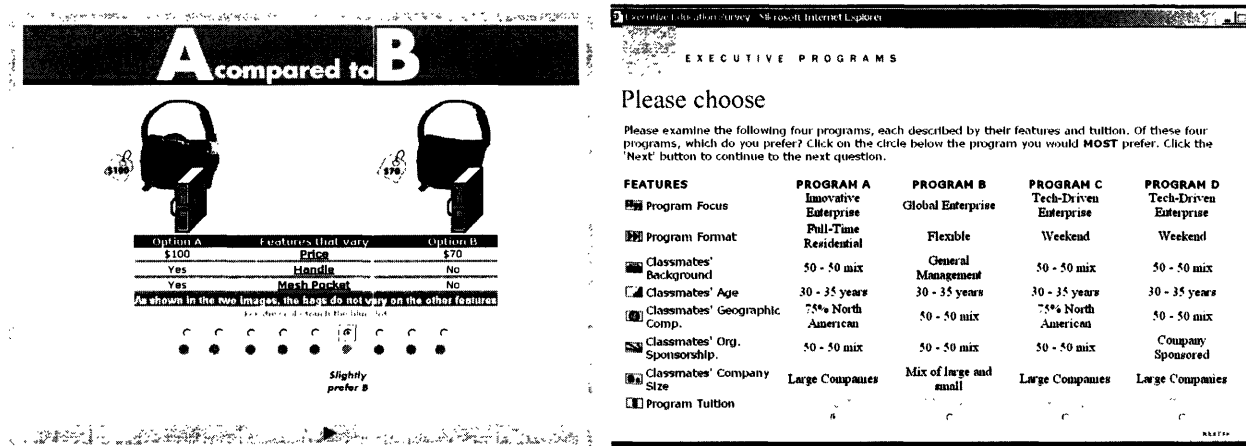
This paper uses stylized models to explore efficient preference questions. The formal analyses provide insights on criteria that are common in widely-used question-design algorithms. We demonstrate that the criterion of metric utility balance leads to inefficient questions, absolute biases, and relative biases among partworths. On the other hand, choice balance for stated-choice questions (the analogy of utility balance for metric questions) can lead to efficiency under some conditions, especially for discrete features. When at least one focal feature is continuous, choices are imbalanced at optimal efficiency. This choice imbalance declines as responses become more accurate and non-focal features increase. The analysis of utility- and choice-balance provide insights on the relative success of metric- and choice-based adaptive polyhedral methods and suggest future improvements.

We then use the formal models to explore a new form of data collection that has become popular in e-commerce and web-based market research – configurators. We study the stylized fact that questions normally used as warm-up questions in conjoint analysis are surprising accurate in predicting consumer behavior that is relevant to e-commerce applications. For these applications, new data suggest that warm-up questions are, perhaps, more useful managerially than paired-comparison tradeoff questions. We examine and generalize this insight to demonstrate the advantages of matching tradeoff, configurator, and “range” questions to the relevant managerial decisions. This further generalizes to M-efficiency (managerial efficiency), an alternative criteria for question design algorithms. Designs which minimize D-errors and A-errors may not minimize M_D - and M_A -errors which measure precision relevant to managerial decisions. We provide examples in which simple modifications to question design increase M-efficiency.

1. Motivation

Preference measurement is crucial to managerial decisions in marketing. Methods such as conjoint analysis, voice-of-the-customer methods, and, more recently, configurators, are used widely for product development, advertising, and strategic marketing. In response to this managerial need, marketing scientists have developed many sophisticated and successful algorithms to select questions efficiently. For example, Figure 1 illustrates two popular formats that have been studied widely.

Figure 1
Popular Formats for Conjoint Analysis



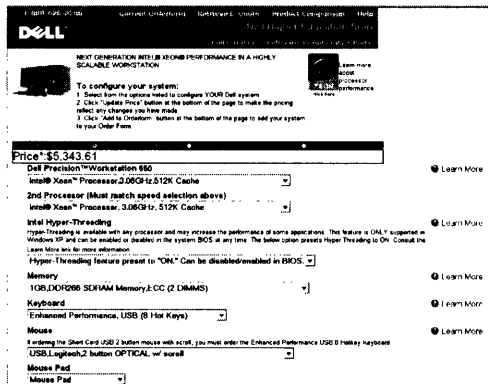
(a) Metric paired-comparison format (laptop bags)

(b) Stated-choice format (Executive Education)

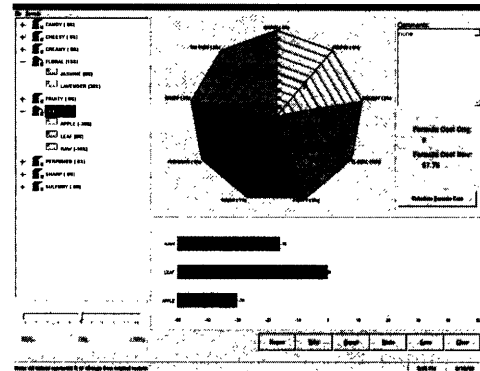
For the metric-pairs format (Figure 1a) questions can be selected with efficient designs, with Adaptive Conjoint Analysis (ACA), or with polyhedral methods (Kuhfield, Tobias, and Garratt 1994; Sawtooth 2002; Toubia, Simester, Hauser, Dahan 2003). For the stated-choice format (Figure 1b), recent advances adapt questions based on pretests, managerial beliefs, or prior questions and do so either by selecting an aggregate design that is replicated for all subsequent respondents or by adapting question design for each respondent (Arora and Huber 2001; Huber and Zwerina 1996; Johnson, Huber and Bacon 2003; Kanninen 2002; Sandor and Wedel 2001, 2002, Toubia, Hauser, and Simester 2003). These question-design methods are used with a variety of estimation methods (regression, Hierarchical Bayes estimation, mixed logit models,

analytic center approximations, and hybrid methods) and have proven to provide more efficient and more accurate data with which to estimate partworths.

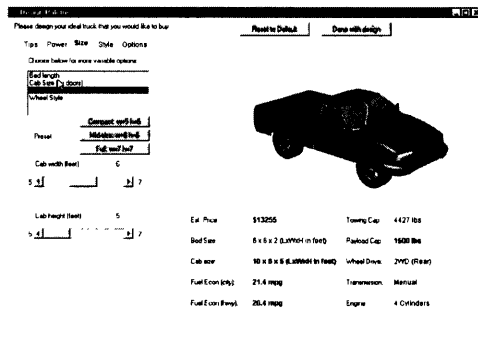
Figure 2
Example Configurators



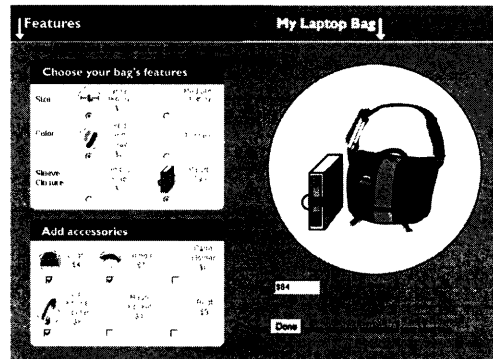
(a) Configurator from Dell.com



(b) Innovation Toolkit for Industrial Flavorings



(c) Design Palette for Pickup Trucks



(d) User Design for Laptop Computer Bags

Recently, as preference measurement has moved to web-based questionnaires, e-commerce firms and academic researchers have experimented with a new format in which respondents (or e-commerce customers) select features to configure a preferred product. The best-known e-commerce example is Dell.com (Figure 2a); however, the format is becoming popular on many websites, especially automotive websites (e.g., kbb.com's auto choice advisor, vehix.com, autoweb.com). For marketing research, configurators are known variously as innovation toolkits, design palettes, and user design formats. They have been instrumental in the design of flavorings for International Flavor and Fragrance, Inc (Figure 2b), new truck platforms for

General Motors (Figure 2c), and new messenger bags for Timbuk2 (Figure 2d). (Figures from Dahan and Hauser 2002; Thomke and von Hippel 2002, Urban and Hauser 2003; von Hippel 2001). Respondents find configurators easy to use and a natural means to express preferences. However, because configurators focus on the respondents' preferred product, they have only been used to estimate partworth estimates when respondents are asked to repeat the tasks with varying stimuli (e.g., Liechty, Ramaswamy and Cohen 2001).

In this paper, we examine question design for all three forms of preference measurement: metric pairs, stated choices, and configurators. We use formal modeling methods to examine key properties that have been used by the algorithms proposed to date. With these models we seek to understand basic properties in order to help explain why some algorithms perform well and others do not. These models also provide insight on the role of configurators in preference measurement and indicate the strengths and weaknesses of this new form of data collection relative to more-standard methods. The models themselves are drawn from empirical experience and represent, in a stylized manner, the basic properties of question design.

The formal models also provide insight on the relationship of question design to managerial use suggesting, for example, that some measurement formats are best for developing a single product, some for product-line decisions, and some for e-commerce decisions. These insights lead to revised managerial criteria (M_A -error and M_D -error).

The paper is structured as follows:

We examine first the concept of utility balance as applied to the metric paired-comparison format and demonstrate that this criterion leads to both bias and inefficiency for that format. This analysis also lends insight on recent comparisons between ACA and polyhedral methods (Toubia, et. al. 2003).

- We then examine a related concept for the stated-choice format. The literature often labels this concept "utility balance," but, for the purpose of this paper, we label it "choice balance" to indicate its relationship to stated-choice data. Our analyses indicate that, unlike utility-balance for metric data, choice balance for stated-choice data can lead to improved efficiency under some conditions. In other cases, optimal efficiency requires unbalanced choices. Our analyses both reflect insights from recent algorithmic papers and provide new insights on how choice balance changes as a function of the characteristics of the problem.

- We then examine data that suggests that configurator questions provide more useful information for e-commerce decisions than “tradeoff” questions (Figure 1). We explore this observation with a stylized model to suggest how question formats are best matched to managerial decisions. These insights generalize to M-efficiency.

We define our scope, in part, by what we do not address. Although we draw experience from and seek to provide insight for new algorithms, algorithmic design is not the focus of this paper. Nor do we wish to enter the debate on the relative merits of metric-rating or stated-choice data. For now we leave that debate to other authors such as Elrod, Louviere and Davey (1992) or Moore (2003). We study both data formats (and configurators) and use formal models to gain insights on all three forms of preference questions. We do not address hierarchical models of respondent heterogeneity, but rather focus on each respondent individually. We recognize the practical importance of hierarchical models, but abstract from such estimation to focus on other basic properties. The recent emergence of individual-specific question designs and/or differentiated designs suggest that our insights should extend to such analyses (Johnson, Huber, and Bacon 2003; Sandor and Wedel 2003; Toubia, Hauser, and Simester 2003). But that remains future research.

2. Efficiency in Question Design

Before we begin our analyses, we review briefly the concepts of efficient question design. The concept of question efficiency has evolved to represent the accuracy with which partworths are estimated. Efficiency focuses on the standard errors of the estimates. Let \vec{p} be the vector of partworths to be estimated and $\hat{\vec{p}}$ the vector of estimated partworths. Then, if $\hat{\vec{p}}$ is (approximately) normally distributed with variance Σ , the confidence region for the estimates is an ellipsoid defined by $(\vec{p} - \hat{\vec{p}})' \Sigma^{-1} (\vec{p} - \hat{\vec{p}})$, Greene (1993, p. 190).³⁷ For most estimation methods, Σ depends upon the questions that the respondent is asked and, hence, an efficient set of questions minimizes the confidence ellipsoid. This is implemented as minimizing a norm of the matrix, Σ . A-errors are based on the trace of Σ , D-errors on the determinant of Σ , and G-errors on the maximum diagonal element. Because the determinant is often the easiest to optimize, most

³⁷ In practice, Σ is unknown and is replaced with its estimated value. For simplicity we assume that Σ is known, which should change neither the insights nor the results of our analyses.

empirical algorithms work with D-efficiency (Kuhfeld, Tobias, and Garratt 1994, p. 547). G-efficiency is rarely used.

Because the determinant of a matrix is the product of its eigenvalues and the trace is the sum of its eigenvalues, A-errors and D-errors are often highly correlated (Kuhfeld, Tobias, and Garratt 1994, p. 547, Arora and Huber 2001, p. 274). In this paper we use both, choosing the criterion which leads to the most transparent insight. In some cases the results are easier to explain if we work directly with Σ^{-1} recognizing $\det(\Sigma) = [\det(\Sigma^{-1})]^{-1}$.³⁸

3. Utility Balance for the Metric Paired-Comparison Format

The metric pairs format (Figure 1a) is widely used for commercial conjoint analysis (Green, Krieger, Wind 2001, p. S66; Ter Hofstede, Kim and Wedel 2002, p. 259). Although the format has been used academically and commercially for over 25 years, its most common application is in the adaptive component of ACA (Green, Krieger and Agarwal 1991, Hauser and Shugan 1980, Johnson 1991).³⁹

Metric utility balance is at the heart of ACA. With the goal that “a difficult choice provides better information for further refining utility estimates,” “ACA presents to the respondent pairs of concepts that are as nearly equal as possible in estimated utility (Orme 1999, p. 2; Sawtooth Software, 2002, p. 11, respectively).” This criterion has appeal. For example, in the study of response latencies, Haaijer, Kamakura, and Wedel (2000, p. 380) suggest that respondents take more time on choice sets that are balanced in utility and, therefore, make less error-prone choices. Green, Krieger and Agarwal (1991, p. 216) quoting a study by Huber and Hansen (1986) ascribe better predictive ability “to greater respondent interest in these difficult-to-judge pairs.” Finally, Shugan (1980) provides a theory to suggest that the cost of thinking is inversely proportional to the square of the utility difference. There are clear advantages in terms of respondent interest, attention, and accuracy for metric utility-balanced questions.

³⁸ The trace of a matrix does not necessarily equal the inverse of the trace of the inverse matrix, thus an empirical algorithm that maximizes the trace of Σ^{-1} does not necessarily minimize the trace of Σ . However, the eigenvalues are related, especially when the eigenvalues of Σ are similar in magnitude, as they are in empirical situations.

³⁹ For example, Dr. Lynd Bacon, AMA's Vice President for Market Research and Executive Vice President of NFO Worldgroup confirms that metric pairs remain one of the most popular forms of conjoint analysis. Sawtooth Software claims that ACA is the most popular form of conjoint analysis in the world, likely has the largest installed base, and, in 2001, accounted for 37% of their sales. Private communications and *Marketing News*, 04/01/02, p. 20.

However, recent simulation and empirical evidence suggests that alternative criteria provide question designs that lead to more accurate partworth estimates (Toubia, et. al. 2003). Furthermore, adaptation implies endogeneity in the design matrix, which, in general, leads to bias. Thus, we explore whether or not the criterion of metric utility balance has adverse characteristics, such as systematic biases, that offset the advantages in respondent attention. We begin with a stylized model.

A Stylized Model

For simplicity we consider products with two features, each of which is specified at two levels. We assume further that the features satisfy preferential independence such that we can write the utility of a profile as an additive sum of the partworths for the features of the profile (Keeney and Raiffa 1976, p. 101-105; Krantz, et. al. 1971). Without loss of generality, we scale the low level of each feature to zero. Let p_1 be the partworth of the high level of feature 1 and let p_2 be the partworth of the high level of feature 2. For these assumptions there are four possible profiles which we write as $\{0, 0\}$, $\{0, 1\}$, $\{1, 0\}$, and $\{1, 1\}$ to indicate the levels of features 1 and 2, respectively. With this notation we have the true utilities given by:

$$u(0, 0) = 0 \qquad u(1, 0) = p_1 \qquad u(0, 1) = p_2 \qquad u(1, 1) = p_1 + p_2$$

For metric questions in which the respondent provides an interval-scaled response, we model response error as an additive, zero-mean random variable, e . For example, if the respondent is asked to compare $\{1, 0\}$ to $\{0, 1\}$, the answer is given by $a_{12} = p_1 - p_2 + e$. We denote the probability distribution of e with $f(e)$. We denote the estimates of the partworths, estimated from responses to the questions, by \hat{p}_1 and \hat{p}_2 .

Adaptive (Metric) Utility Balance

Without loss of generality, consider the case where $2p_2 > p_1 > p_2 > 0$ such that the off-diagonal question ($\{0,1\}$ vs. $\{1,0\}$) is the most utility-balanced question. We assume that this ordinal relationship is based on prior questions. We label the error associated with the most utility-balanced question as e_{ub} and label errors associated with subsequent questions as either e_1 or e_2 . We now consider an algorithm which adapts questions based on utility balance. Adaptive metric utility balance implies the following sequence.

First question: $\hat{p}_1 - \hat{p}_2 = p_1 - p_2 + e_{ub}$

Second question: $\hat{p}_1 = p_1 + e_1$ if $e_{ub} < p_2 - p_1$ (Case 1)

$\hat{p}_2 = p_2 + e_2$ if $e_{ub} \geq p_2 - p_1$ (Case 2)

Suppose that $e_{ub} \geq p_2 - p_1$, then:

$$E[\hat{p}_2] = p_2 + E[e_2] = p_2$$

$$E[\hat{p}_1] = E[\hat{p}_2] + p_1 - p_2 + E[e_{ub} | e_{ub} \geq p_2 - p_1] = p_1 + E[e_{ub} | e_{ub} \geq p_2 - p_1] > p_1$$

Suppose that $e_{ub} < p_2 - p_1$, then:

$$E[\hat{p}_1] = p_1 + E[e_1] = p_1$$

$$E[\hat{p}_2] = E[\hat{p}_1] + p_2 - p_1 - E[e_{ub} | e_{ub} < p_2 - p_1] = p_2 - E[e_{ub} | e_{ub} < p_2 - p_1] > p_2$$

Thus, under both Case 1 and Case 2, one of the partworths is biased upwards for any zero-mean distribution that has non-zero measure below $p_2 - p_1$.

The intuition underlying this proof is that the second question in the adaptive algorithm depends directly on the random error – a direct version of endogeneity bias (Judge, et. al. 1985, p. 571). We demonstrate how to generalize this result with a pseudo-regression format of

$\hat{\vec{p}} = (X'X)^{-1}(X'\vec{a}) = \vec{p} + (X'X)^{-1}X'\vec{e}$, where \vec{a} and \vec{e} are the answer and error vectors and X is the question matrix. Then the second row of X depends on the magnitude of e_{ub} relative to $p_2 - p_1$. Specifically, the first row of X is $[1, -1]$ and its second row can be written as $[b, 1-b]$ where $b = 1$ if errors are small and $b = 0$ if errors are large. The bias, \vec{B} , is then given by the following equation where $k = 1$ or 2 .

$$(1) \quad E[\vec{B}] = E[(X'X)^{-1}X'e] = E\left(\begin{bmatrix} (1-b)e_{ub} + e_k \\ -be_{ub} + e_k \end{bmatrix}\right) = E\left(\begin{bmatrix} (1-b)e_{ub} \\ -be_{ub} \end{bmatrix}\right)$$

Equation 1 illustrates the phenomenon clearly: $b = 1$ if e_{ub} is small (more negative) and $b = 0$ if e_{ub} is large (more positive), hence either \hat{p}_1 or \hat{p}_2 is biased upward. These arguments generalize readily to more than two features and more than two questions. The basic insight is that the errors from earlier questions determine the later questions (X) in a systematic way – the essence of endogeneity bias.

Relative Bias

Bias alone does not necessarily impact managerial decisions because partworths are unique only to a positive linear transformation. Thus, we need to establish that the bias is also relative, i.e., the bias depends on the magnitude of the true partworth. Because the bias is proportional to e_{ub} we need to show that $\frac{\partial}{\partial p_1} E[e_{ub} | e_{ub} \geq p_2 - p_1] < 0$ to demonstrate that bias is greater for smaller (true) partworths than it is for larger (true) partworths. We do so by direct differentiation in the Appendix. Thus, we can state the following proposition.

Proposition 1. For the stylized model, on average, metric utility balance biases partworths upward and does so differentially depending upon the true values of the partworths.

When we expand the adaptive questions to a more complex world, in which there are more than two features, the intuition still holds; however, the magnitude of the bias depends upon the specific manner by which metric utility balance is achieved and on the accuracy of the respondent's answers. As an illustration, we built a 16-question simulator for a two-partworth problem in which, after the first question, all subsequent questions were chosen to achieve metric utility balance based on partworths estimated from prior questions. Based on 6,000 regressions each, biases ranged from 1% for relatively large true partworths to 31% for relatively small partworths.⁴⁰ To examine the practical significance we cite data from Toubia, et. al. (2003) who simulate ACA question design for a 10-partworth problem. In their data, biases were approximately 6.6% of the magnitudes of the partworths at eight questions when averaged across the conditions in their experimental design. The biases are significant at the 0.05 level.

The Effect of Metric Utility Balance on Efficiency

Although metric utility balance leads to bias, it might be the case that adaptation leads to greater efficiency. For D-errors, minimizing the determinant of Σ is the same as maximizing the determinant of Σ^{-1} . For metric data, $\Sigma^{-1} = X'X$ (for appropriately centered X). From this ex-

⁴⁰ The simulator is available from <http://mitsloan.mit.edu/vc>. The simulator is designed to demonstrate the phenomenon; we rely on Toubia, et. al. (2003) for an estimate of the empirical magnitude. The simulator chooses x_1 uniformly on $[0,1]$ and x_2 based on utility balance. The absolute values of the response errors are approximately half those of the dependent measure. The magnitude of the bias changes if the response errors change, but the direction of the bias remains.

pression we see two things immediately. First, as Kanninen (2001, p. 217) and Kuhfeld, et. al. (1994, p. 547) note, for a given set of levels, efficiency tends to push features to their extreme values. (These statements assume that the maximum levels are selected to be as large as is reasonable for the empirical problem. Levels which are too extreme risk additional respondent error not explicitly acknowledged by the statistical model used in question design. See Delquie 2003). Second, perfect metric utility balance imposes a linear constraint on the columns of X because $X\bar{p} = \vec{0}$. This constraint induces $X'X$ to be singular. When $X'X$ is singular, the determinant will be zero and D-errors increase without bound. Imperfect metric utility balance leads to $\det(X'X) \approx 0$ and extremely large D-errors. A-errors behave similarly. Thus, greater metric utility balance tends to make questions inefficient.⁴¹

Insights on the Relative Performance of Metric Polyhedral Methods

The analysis of metric utility balance provides insights on the relative performance of metric polyhedral methods. Metric polyhedral methods focus on asking questions to reduce the feasible set of partworths as rapidly as possible by solving an eigenvalue problem related to Sonnevend ellipsoids. These ellipsoids focus questions relative to the “axis” about which there is the most uncertainty. Polyhedral methods do not attempt to impose metric utility balance. In algorithms to date, polyhedral methods are used for (at most) the first n questions where n equals the number of partworths. However, metric polyhedral methods also impose the constraint that the rows of X be orthogonal (Toubia, et. al. 2003, Equation A9). Thus, after n questions, X will be square, non-singular, and orthogonal ($XX' \propto I$ implies $X'X \propto I$).⁴² Subject to scaling, this orthogonality relationship minimizes D-error (and A-error). Thus, while the adaptation inherent in metric polyhedral methods leads to endogenous question design, the orthogonality constraint appears to overcome the tendency of endogeneity to increase errors. Toubia, et. al. (2003) report at most a 1% bias for metric polyhedral question selection, significantly less than observed for ACA question selection.

⁴¹ The simulator cited in the acknowledgements section of this paper also contains an example in which metric utility balance can be imposed. The more closely utilities are balanced, the worse the fit.

⁴² Orthogonal rows assume that $XX' = \alpha I$ where α is a proportionality constant. Because X is non-singular, X and X' are invertible, thus $X'XX'X'^{-1} = \alpha X'X'^{-1} \Rightarrow X'X = \alpha I$.

Summary

Metric utility balance may retain respondent interest and encourage respondents to think hard about tradeoffs, however, these benefits come at a price. Metric utility balance leads to bias, relative bias, and lowered efficiency. On the other hand, the inherent orthogonality criterion (rather than utility balance) in metric polyhedral methods enables metric polyhedral question design to achieve the benefits of adaptation with significantly less bias and, apparently, without reduced efficiency.

4. Choice Balance for the Stated-Choice Format

In the stated-choice format (Figure 1b), two profiles that have the same utility will be equally likely to be chosen, thus choice balance is related to utility balance. However, because the underlying statistical models are fundamentally different, the implications of choice balance for stated-choice questions are different from the implications of metric utility for metric paired-comparison questions.

Choice balance is an important criterion in many of the algorithms used to customize stated-choice questions. For example, Huber and Zwerina (1996) begin with orthogonal, level-balanced, minimal overlap designs and demonstrate that swapping and relabeling these designs for greater choice balance increases D_p -efficiency. Arora and Huber (2001, p. 275) test these algorithms in a variety of domains with hierarchical Bayes methods commenting that the maximum D_p -efficiency is achieved by sacrificing orthogonality for better choice balance.⁴³ Orme and Huber (2001, p. 18) further advocate choice balance. Toubia, Hauser, and Simester (2003) use choice-balance to increase accuracy in adaptive choice-based questions. Finally, the concept of choice balance is germane to the principle, used in the computer adaptive testing literature, that a test is most effective when its items are neither too difficult nor too easy, such that the probabilities of correct responses are close to 0.50 (Lord 1970).

To understand better why choice balance improves efficiency while metric utility balance degrades efficiency, we recognize that, for logit estimation of partworths with stated-choice data, Σ^{-1} depends upon the true partworths and, implicitly, on response error. Specifically, McFadden (1974) shows that:

⁴³ D_p -efficiency recognizes that Σ depends on the true partworths. See Equation 2.

$$(2) \quad \Sigma^{-1} = R \sum_{i=1}^q \sum_{j=1}^{J_i} (\bar{x}_{ij} - \sum_{k=1}^{J_i} \bar{x}_{ik} P_{ik})' P_{ij} (\bar{x}_{ij} - \sum_{k=1}^{J_i} \bar{x}_{ik} P_{ik})$$

where R is the effective number of replicates; J_i is the number of profiles in choice set i ; q is the number of choice sets; \bar{x}_{ij} is a row vector describing the j^{th} profile in the i^{th} choice set; and P_{ij} is the probability that the respondent chooses profile j from the i^{th} choice set.

Equation 2 has a more transparent form when all choice sets are binary, in particular:

$$(3) \quad \Sigma^{-1} = R \sum_{i=1}^q (\bar{x}_{i1} - \bar{x}_{i2})' P_{i1} (1 - P_{i1}) (\bar{x}_{i1} - \bar{x}_{i2}) = R \sum_{i=1}^q \vec{d}_i' P_{i1} (1 - P_{i1}) \vec{d}_i$$

where \vec{d}_i is the row vector of differences in the features for the i^{th} choice set. In Equation 3 it is clear that choice balance ($P_{i1} \rightarrow 1/2$) tends to increase a norm of Σ^{-1} . However, the choice probabilities (P_{i1}) are a function of the feature differences (\vec{d}_i), thus pure choice balance may not maximize either the determinant (or the trace) of Σ^{-1} .

First recognize that the true choice probabilities in the logit model are functions of the utilities, $\vec{d}_i \vec{p}$. If we hold the utilities constant, we also hold the choice probabilities constant. Subject to this linear utility constraint, we can increase efficiency without bound by increasing the absolute magnitudes of the \vec{d}_i 's. This result is similar to that for the metric-pairs format; the experimenter should choose feature differences that are as large as reasonable subject to the linear constraint imposed by $\vec{d}_i \vec{p}$ and subject to any errors (outside the model) introduced by large feature differences (Delquie 2003).

Kanninen (2002) exploits these properties effectively by selecting one continuous feature (e.g. price) as a focal feature with which to manipulate $\vec{d}_i \vec{p}$, while treating all remaining features as discrete by setting them to their maximum or minimum levels. She reports substantial increases in efficiency for both binary and multinomial choice sets (p. 223).

Kanninen uses numerical means to select the optimal levels of the continuous focal feature. We gain further insight with formal analysis. In particular, we examine the optimality conditions for the focal feature, holding the others constant. For greater transparency we use bi-

nary questions and the trace of Σ^{-1} rather than the determinant.⁴⁴ The trace is given by:

$$(4) \quad \text{trace}(\Sigma^{-1}) = \sum_{i=1}^q \sum_{k=1}^K d_{ik}^2 P_{i1} (1 - P_{i1})$$

If we assume, without loss of generality, that the focal feature is the K^{th} feature, then the first-order conditions for the focal feature are:

$$(5) \quad d_{iK} = \left(p_K \sum_{k=1}^K d_{ik}^2 \right) \left(\frac{P_{i1} - P_{i2}}{2} \right) \text{ for all } i$$

where d_{ik} is the level of the k^{th} feature difference in the i^{th} binary question.

We first rule out the trivial solution of $d_{iK} = 0$ for all k . Not only would this imply a choice between two identical alternatives, but the second-order conditions imply a minimum. A slightly more interesting case occurs when perfect choice balance is already achieved by the $K-1$ discrete features. We return to this solution later, however, in the focal-feature case it occurs with zero probability (i.e., on a set of zero measure). The more interesting solutions for a continuous focal feature require that d_{iK} be non-zero and, hence, that $P_{i1} \neq P_{i2}$. Thus, when d_{iK} can vary continuously, Equation 5 implies that optimal efficiency requires choice imbalance.

There are two equivalent solutions to Equation 5 corresponding to a right-to-left switch of the two alternatives. That is, the sign of d_{iK} must match the sign of $P_{i1} - P_{i2}$. Thus, without loss of generality, we examine situations where d_{iK} is positive (see Appendix). We plot the marginal impact of d_{iK} on $\text{trace}(\Sigma^{-1})$ for illustrative levels of the non-focal feature difference.⁴⁵ Following Arora and Huber (2001) and Toubia, Hauser and Simester (2003), we use the magnitude of the partworths as a measure of response accuracy – higher magnitudes indicate higher response accuracy. Figure 3 illustrates that efficiency is maximized for non-zero focal feature differences. Figure 3 also illustrates that the optimal level of the focal feature difference depends upon the level of the non-focal features and upon the magnitude.

Choice balance ($P_{i1} - P_{i2}$) is a non-linear function of the feature differences. Nonetheless, we can identify the systematic impact of both non-focal features and response accuracy. In

⁴⁴ For a focal feature, the trace and the determinant have related maxima, especially if all but the non-focal feature are set to their minimum or maximum values.

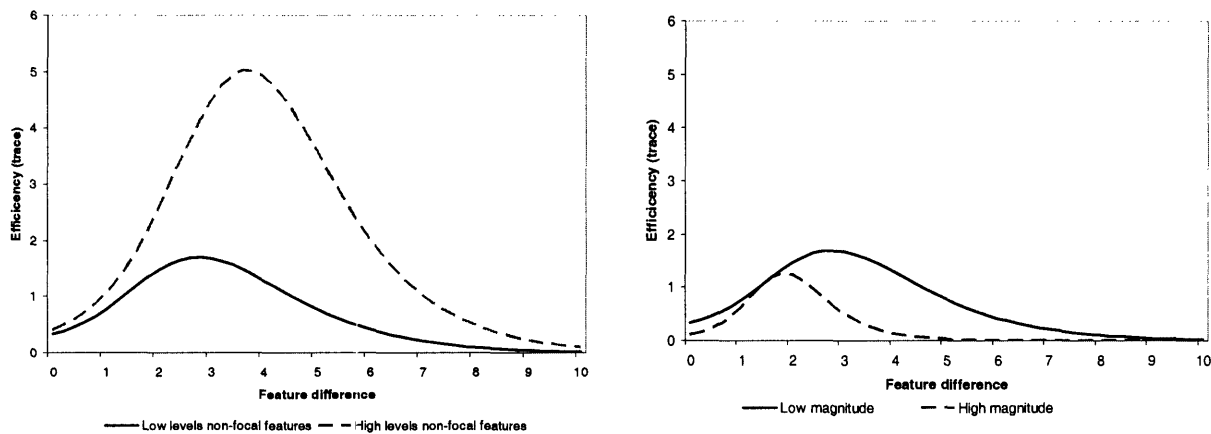
⁴⁵ For low magnitudes, the partworths of both feature are set equal to 1.0 in these plots. For high magnitudes they are doubled. The low level of the non-focal feature (absolute value) is 1.5. The high level is 3.0.

particular, formal analyses reveal that choice balance increases with greater response accuracy and with greater levels of the non-focal feature differences. We formalize these insights in two propositions. The proofs are in the Appendix.

Proposition 2. As response accuracy increases, choice balance increases and the level difference in the focal feature decreases.

Proposition 3. As the levels of the non-focal features increase, choice balance increases.

Figure 3
Optimal Efficiency and Choice Imbalance



(a) Low vs. high non-focal feature levels

(b) Low vs. high magnitude in logit model

Quantal Features

In many applications, all features are binary – they are either present or absent. Examples include four-wheel steering on automobiles, cell-phone holders on laptop bags, and auto focus on cameras. In these cases, the focal-feature method cannot be used for maximizing efficiency. However, discrete-search algorithms are feasible to maximize a norm on Σ^{-1} . Arora and Huber (2001) and Sandor and Wedel (2001) provide two such algorithms. Sandor and Wedel (2002) extend these concepts to a mixed logit model.

An interesting (and common) case occurs when the experimenter limits the number of features that vary, perhaps due to cognitive load (Sawtooth Software 1996, p. 7; Shugan 1980).

For a given number of binary features (± 1), $\sum_k d_{ik}^2$ is fixed. In this case, it is easy to show that Equation 4 is maximized when choices are as balanced as feasible.⁴⁶ This same phenomenon applies when the researcher requires that all features vary.

Insights on the Relative Performance of Choice-based Polyhedral Methods

Like metric polyhedral methods, choice-based polyhedral methods rely upon Sonnevend ellipsoids. However, for stated-choice data, respondents' answers provide inequality constraints rather than equality constraints. Rather than selecting questions vectors parallel to the longest axis of the ellipsoid, choice-based polyhedral methods use the ellipsoid to identify target part-worths at the extremes of the feasible region. The choice-based algorithm then solves a utility maximization (knapsack) problem to identify the features of the profiles. The knapsack problem assures that the constraints go (approximately) through the analytic center of the region. This criterion provides approximate a priori choice balance. (By a priori we mean that, given the prior distribution of feasible points, the predicted choices are equal.) In this manner, the algorithm achieves choice balance without requiring utility balance throughout the region. Utility balance only holds for the analytic center of the region – a set of zero measure.

The applications of choice-based polyhedral methods to date have focused on discrete features implemented as binary features. For discrete features, choice balance achieves high efficiency. For continuous features, Proposition 2 suggests algorithmic improvements. Specifically, the knapsack problem might be modified to include $trace(\Sigma^{-1})$ in the objective function. However, unlike a knapsack solution, which is obtained rapidly, the revised math program may need to rely on new, creative heuristics.

Summary

Choice balance affects the efficiency of stated-choice questions quite differently than metric utility balance affects metric paired-comparison questions. This, despite the fact that choice balance is related to ordinal utility balance. If at least one feature is continuous, then that feature can be used as a focal feature and the most efficient questions require choice imbalance

⁴⁶ If there are a maximum number of features that can be non-zero rather than a fixed number, it is possible, although less likely, that a maximum will occur for less than the maximum number of features. This case can be checked numerically for any empirical problem. Nonetheless, the qualitative insight holds.

(and non-zero levels of the focal feature). However, we show formally that more accurate responses and higher non-focal features lead to greater choice balance. If features are discrete and the number which can vary is fixed, then choices should be as balanced as feasible. Finally, this analysis helps explain the relative performance of stated-choice polyhedral methods and aggregate customization methods.

5. Configurators

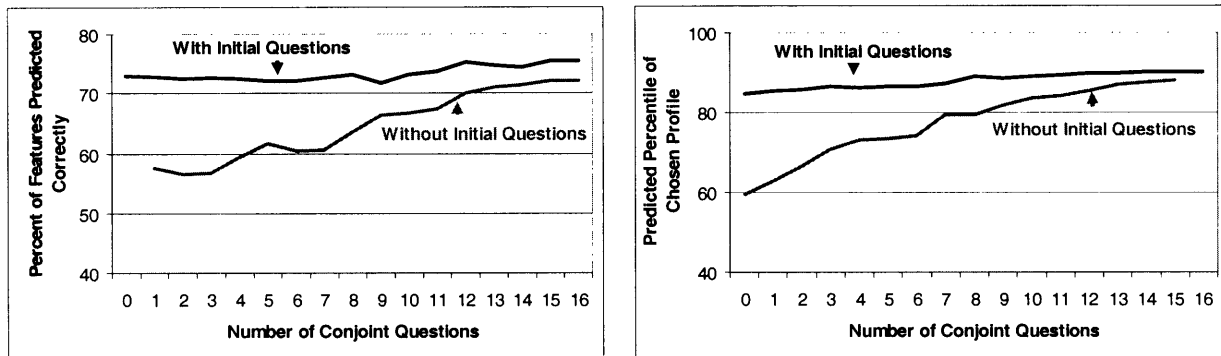
In Sections 3 and 4 we explored two popular formats for preference questions suggesting that, for the metric paired-comparison format, utility balance led to bias, relative bias and inefficiency while, for the choice-based format, choice balance could be efficient for discrete features but not necessarily for continuous focal features. We now explore a relatively new format, configurators, in which respondents focus on the features one at a time and are not asked to compare profiles in the traditional sense. Neither utility balance nor choice balance apply to configurators. Nonetheless, analysis of this new format leads to insights that generalize to all three question formats and helps us understand which format is most efficient in which situations.

In configurators, respondents are typically given a price for each feature level and asked to indicate their preferences for that price/feature combination (review Figure 2). Configurator questions are equivalent to asking the respondent whether the partworths of the features (p_k 's) are greater than or less than the stated prices. These data appear quite useful. For example, Liechty, Ramaswamy and Cohen (2001) use ten repeated configurator measures per respondent to estimate successfully hierarchical Bayes probit models.

Stylized Fact: Warm-up Questions do Surprisingly Well in Predicting Feature Choice

Figure 4 presents an empirical example. In data collected by Toubia, et. al. (2003), respondents were shown five laptop bags worth approximately \$100 and allowed to choose among those bags. Various conjoint methods based on metric paired-comparison questions collected prior to this task predicted these profile choices quite well.

Figure 4
Warm-up Questions are Effective Predictors of Configurator Tasks



(a) Ability to predict feature choice

(b) Ability to predict preferred profile

Before completing the metric paired-comparison task, in order to familiarize respondents with the features, respondents answered warm-up questions. They were asked to indicate the preferred level of each feature. Such warm-up questions are common. We re-analyzed Toubia, et. al.'s data in two ways. One conjoint estimation method did not use the warm-up questions. The other conjoint estimation method included monotonic constraints based on the warm-up questions.⁴⁷ One month after the initial conjoint-analysis questionnaire, the respondents were recontacted and given a chance to customize their chosen bag with a configurator. Figure 4 plots the ability of the conjoint analyses to predict configurator choice as a function of the number of metric paired-comparison questions that were answered.

For both predictive criteria, the ability to predict configurator choice is surprising flat when initial warm-up questions are included. (Predictions are less flat for forecasts of the choice of five laptop bags, Toubia, et. al. 2003, Table 6.) For configurator predictions, the nine warm-up questions appear to provide more usable information than the sixteen paired-comparison questions. This empirical observation is related to observations by Evgeniou, Boussios, and Zacharia (2003) whose simulations suggest the importance of ordinal constraints. While, at first, Figure 4 appears to be counter-intuitive, it becomes more intuitive when we realize that the

⁴⁷ In the method labeled “with initial questions,” the warm-up questions were used to determine which of two feature levels were preferred (e.g., red or black). In the method labeled “without initial questions,” such ordering data were not available to the estimation algorithm. In this particular plot, for consistency, we use the analytic-center method of Toubia, et. al., however, the basic insights are not limited to this estimation method.

warm-up questions are similar to configurator choice and the paired-comparison questions are similar to the choice among five (Pareto) profiles. We formalize this intuition with our stylized model.

Stylized Model to Explain the Efficiency of Warm-up Questions

For transparency assume metric warm-up questions. This enables us to compare the warm-up questions to paired-comparison questions without confounding a metric-vs.-ordinal characterization. We consider the same three questions used previously, rewriting e_{ub} as e_i for mnemonic exposition of paired-comparison tradeoff questions. Without loss of generality we incorporate feature prices into the definition of each feature. We assume $p_1 > p_2 > 0$, but need not assume $2p_2 > p_1$ as we did before. The three questions are:

- (Q1) $\hat{p}_1 = p_1 + e_1$
- (Q2) $\hat{p}_2 = p_2 + e_2$
- (Q3) $\hat{p}_1 - \hat{p}_2 = p_1 - p_2 + e_i$

We now consider stylized estimates obtained from two questions. In our stylized world of two features at two levels, Q1 and Q2 can be either metric warm-up questions or metric paired-comparison questions. Q3 can only be a metric paired-comparison question. Assuming all questions are unbiased, we examine which combination of questions provides the greatest precision (efficiency): Q1+ Q2, Q1+Q3, or Q2+Q3.

Consider feature 1. Because feature price is incorporated in the feature, if $p_k \geq 0$, then feature choice is correctly predicted if \hat{p}_k is non-negative. (We use similar reasoning for feature 2.) Simple calculations (see Appendix) reveal that the most accurate combination of two questions is Q1+Q2 if and only if $\text{Prob}[e_k \geq -p_k] > \text{Prob}[e_k + e_i \geq -p_k]$. Since $-p_k$ is a negative number, these conditions hold for most reasonable density functions. For example, in the Appendix we demonstrate that, if the errors are independent and identically distributed normal random variables, then this condition holds.

Proposition 4. For zero-mean normally distributed (i.i.d.) errors, the warm-up questions (Q1+Q2) provide more accurate estimates of feature choice than do pairs of questions that include a metric paired-comparison question (Q3).

Intuitively, metric warm-up questions parse the errors more effectively than tradeoff questions when the focus is on feature choice rather than holistic products. The questions focus precision where it is needed for the managerial decisions. We provide a more formal generalization of these concepts in the next section. The intuition does not depend upon the metric properties of the warm-up questions and applies equally well to ordinal questions. We invite the reader to extend Proposition 4 to ordinal questions using the M-efficiency framework that is explored in the next section.

6. Generalizations of Question Focus: M-Efficiency

The intuition drawn from Proposition 4 suggests that it is best to match preference questions carefully to the managerial task. For example, metric warm-up questions are similar to configurator questions and may be the most efficient for predicting feature choice. On the other hand, tradeoff questions are similar to the choice of a single product from a Pareto set. Estimating partworths from tradeoff questions may be best if the managerial goal is to predict consumer choice when a single new product is launched into an existing market. We now examine this more general question. We begin with the stylized model for two questions and then suggest a generalization to an arbitrary number of questions.

Stylized Analysis: Matching Question Selection to Managerial Focus

With metric questions (and preferential independence) we can ask one more independent question, which we call a “range” question.⁴⁸ We then consider three stylized product-development decisions.

$$(Q4) \quad \hat{p}_1 + \hat{p}_2 = p_1 + p_2 + e_r$$

Launch a single product

The product-development manager has one shot at the market and has to choose the single best product to launch, perhaps because of shelf-space or other limitations. The manager is most interested in the difficult tradeoffs, especially if the two features are substitutes. The manager seeks precision on $p_1 - p_2$.

Launch a product line (platform decision)

⁴⁸ Such questions are feasible with a metric format. They would be eliminated as dominated alternatives in an ordinal format such as stated-choice questions.

The product-development manager is working on a platform from which to launch multiple products, perhaps because of synergies in production. The manager is most interested in the range of products to offer (the breadth of the platform). The manager seeks precision on $p_1 + p_2$.

Launch a configurator-based e-commerce website

The product-development manager plans to sell products with a web-based configurator, perhaps because of flexible manufacturing capability. However, there is a cost to maintaining a broad inventory and there is a cognitive load on the consumer based on the complexity of the configurator. The manager seeks precision on p_1 and on p_2 .

For these stylized decisions we expect intuitively that we gain the most precision by matching the questions to the managerial decisions. Table 1 illustrates this intuition. For decisions on a single product launch, we gain the greatest precision by including tradeoff questions (Q3); for decisions on product lines we gain the greatest precision by including range questions (Q4), and for decisions on e-commerce we gain the greatest precision by focusing on feature-specific questions (Q1 and Q2). These analytical results, which match question format to managerial need (for metric data), are comparable in spirit to the empirical results of Marshall and Bradlow (2002) who examine congruency between the formats of the profile and validation data (metric vs. ranking vs. choice) in hybrid conjoint analysis. Marshall and Bradlow’s analyses suggest that profile data require less augmentation with self-explicated data when they are congruent with the validation data.

Table 1
Error Variances Due to Alternative Question Selection

	Tradeoff Questions	Range Questions	Feature Focus
Single-product launch	σ^2	$3\sigma^2$	$2\sigma^2$
Product-line decisions	$3\sigma^2$	σ^2	$2\sigma^2$
e-Commerce website	$\frac{3}{2}\sigma^2$	$\frac{3}{2}\sigma^2$	σ^2

M-Efficiency

The insights from Table 1 are simple, yet powerful, and generalize readily. In particular, the stylized managerial decisions seek precision on $M\bar{p}$ where M is a linear transform of \bar{p} . For example, if the manager is designing a configurator and wants precision on the valuation of features relative to price, then the manager wants to concentrate precision not on the partworths per se, but on specific combinations of the partworths. We indicate this managerial focus with a matrix, M_c , where the last column corresponds to the price partworth. This focus matrix provides criteria, analogous to D-errors and A-errors, by which we can design questions.

$$M_c = \begin{bmatrix} 1 & 0 & \dots & -1 \\ 0 & 1 & \dots & -1 \\ & & \dots & \\ 0 & \dots & 1 & -1 \end{bmatrix}$$

M-Efficiency for Metric Data

We begin with metric data to illustrate the criteria. For metric data, the standard error of the estimate of $M\bar{p}$ is proportional to $\Sigma_M = M(X'X)^{-1}M'$ (Judge, et. al. 1985, p. 57). By defining efficiency based on Σ_M rather than Σ , we focus precision on the partworths that matter most to the manager's decision. We call this focus managerial efficiency (M-efficiency) and define M-errors as follows for suitably scaled X matrices.

$$M_D\text{-error} = q \det(M(X'X)^{-1}M')^{1/n}$$

$$M_A\text{-error} = q \text{trace}(M(X'X)^{-1}M') / n$$

The managerial focus affects D-errors and A-errors differently. M_D -error is defined by the determinant of the modified covariance matrix and, hence, focuses on the geometric mean of the estimation variances (eigenvalues). All else equal, the geometric mean is minimized if the estimation variances are equal in magnitude. On the other hand, M_A -error is defined by the trace of the covariance matrix and, hence, focuses on the arithmetic mean of the estimation variances. Thus, if the manager wants differential focus on some combinations of partworths and is willing to accept less precision on other combinations, then M_A -errors might be a better metric than M_D -errors.

To illustrate M-efficiency, we provide an example in which we accept higher A- and D-errors in order to reduce M_A - and M_D -errors. Consider an orthogonal array, X . For this design, $X'X = 12I_6$, where I_6 is a 6x6 identify matrix. The orthogonal array minimizes both D-errors and A-errors. It does not necessarily minimize either M_D - or M_A -errors as we illustrate with the five-feature configurator matrix, M_c .

$$M_c = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

$$X = \begin{bmatrix} +1 & -1 & -1 & -1 & -1 & -1 \\ -1 & +1 & -1 & +1 & +1 & +1 \\ +1 & +1 & +1 & -1 & +1 & +1 \\ -1 & -1 & +1 & +1 & -1 & +1 \\ +1 & +1 & -1 & +1 & +1 & -1 \\ +1 & -1 & +1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 & -1 & +1 \\ -1 & -1 & -1 & -1 & +1 & -1 \\ -1 & +1 & -1 & -1 & -1 & +1 \\ -1 & +1 & +1 & -1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 \\ -1 & -1 & +1 & +1 & +1 & -1 \end{bmatrix}$$

To increase M-efficiency, we perturb $X'_M X_M$ closer to $M'_c M_c$ while maintaining full rank. We choose X_M such that $X'_M X_M = \alpha[\beta I_6 + (1 - \beta)M'_c M_c]$ where α and β are scalars, $\beta \in [0, 1]$, and α is chosen to maintain constant question “magnitudes.”⁴⁹ By construction, X_M is no longer an orthogonal array, hence both A-errors and D-errors will increase. However, M_A - and M_D -errors decrease because X_M is more focused on the managerial problem. For exam-

⁴⁹ Recall that efficiency criteria tend to push feature levels to the extremes of the feasible space. Thus, to compare different designs we must attempt to keep the “magnitudes” of the features constant. In this example, we implement that constraint by keeping the trace of $X'_M X_M$ equal to the trace of $X'X$. Our goal is to provide an example where M-efficiency matters. We can also create examples for other constraints – the details would be different, but the basic insights remain.

ple, with $\beta = 4/5$, M_A -errors decrease by over 20%; M_D -errors decrease, but not as dramatically. Larger β 's turn out to be better at reducing M_D -errors, but less effective at reducing M_A -errors.

Decreases in M_A -efficiency are also possible for square M matrices. However, we cannot improve M_D -errors for square M matrices, because, for square M , the question design that minimizes D-error will also minimize M_D -error. This is true by the properties of the determinant: $\det(\Sigma_M) = \det(M(X'X)^{-1}M') = \det(M)\det((X'X)^{-1})\det(M') = \det(M')\det(M)\det((X'X)^{-1}) = \det(M'M)\cdot\det((X'X)^{-1})$. As an illustration, suppose the manager wants to focus on configurator prices (as in M_c), but also wants to focus on the highest price posted in the configurator. To capture this focus, we make M square by adding another row to M_c , $[1 \ 1 \ 1 \ 1 \ 1 \ -1]$. To decrease M_A -error we unbalance X to make price more salient in the question design. For example, a 40% unbalancing decreases M_A -errors by 5% with only a slight increase in M_D -errors (<1%).

We chose these examples to be simple, yet illustrative. We can create larger examples where the decreases in M-errors are much larger. Our perturbation algorithms are simple; more complex algorithms might reduce M-errors further. M-efficiency appears to be an interesting and relevant criterion on which to focus future algorithmic development.

M-Efficiency for Stated-Choice Questions

The same arguments apply to stated-choice questions. The M-efficient criteria become either the determinant (M_D -error) or the trace (M_A -error) of $M(Z'PZ)^{-1}M'$ where Z is a probability-balanced X matrix and P is a diagonal matrix of choice probabilities. See Arora and Huber (2001, p. 274) for notation. As in Section 4, there will be a tradeoff between choice balance and the levels of focal features. For M-errors, that tradeoff will be defined on combinations of the features rather than the features themselves.

Summary of Sections 5 and 6

Configurator questions, based on analogies to e-commerce websites, are a relatively recent form of preference questions. Empirical data suggest that such questions are surprisingly good at predicting feature choice, in large part because they focus precision on managerial issues that are important for e-commerce and mass customization. This insight generalizes. Intuitively, precision (efficiency) is greatest when the question format is matched to the managerial decision.

M_A -errors (and M_D -errors) provide criteria by which match format to managerial relevance. These criteria can be used to modify question-design algorithms.

There remain the challenges of new algorithmic development. Existing algorithms have been optimized and tested for D- and D_p -efficiency. However, Figure 4, Table 1, and the concept of M-errors have the potential to refocus these algorithms effectively.

7. Summary and Future Research

In the last few years great strides have been made in developing algorithms to select questions for both metric paired-comparison preference questions and for stated-choice questions. Algorithms have been developed for a wide range of estimation methods including regression, logit, probit, mixed logit, and hierarchical Bayes (both metric and choice-based). In some cases the algorithms focus on static designs and in other cases on adaptive designs. This existing research has led to preference questions that are more efficient and more accurate as evidenced by both simulation and empirical experiments.

Summary

In this paper, we use stylized models to gain insight on the properties of algorithms and the criteria they optimize. The models abstract from the complexity of empirical estimation to examine properties transparently. In key cases, the stylized models provide insight on empirical facts. While some of the following insights formalize that which was known from prior algorithmic research, many insights are new and provide a basis for further development:

- Metric utility balance leads to relative biases in partworth estimates.
- Metric utility balance is inefficient.
- Choice balance, while related to utility balance, affects stated-choice questions differently than utility balance affects metric questions.
- When features are discrete and the number of features which can vary is fixed, choice balance is a component of optimal efficiency.
- Otherwise, optimal efficiency implies non-zero choice balance.
- As response accuracy and/or the levels of non-focal features increase, optimal efficiency implies greater choice balance.

- Warm-up (configurator) questions are surprisingly accurate for predicting feature choice because they focus precision on the feature-choice decision.
- This concept generalizes. Tradeoff, range, or configurator questions provide greater efficiency when they are matched to the appropriate managerial problem.
- M-errors provide criteria by which both metric and stated-choice questions can be focused on the managerial decisions that will be made based on the data.
- Simple algorithms (perturbation and/or unbalancing) can decrease M-errors effectively.

The formal models also provide a theory with which to understand the accuracy of recently proposed polyhedral methods. Specifically,

- Metric polyhedral methods appear to avoid the bias and inefficiency of utility balance by maintaining orthogonality while focusing questions where partworths estimates are less precise.
- To date, stated-choice polyhedral methods have been applied for discrete features. Hence, their use of approximate choice balance makes them close to efficient.
- However, stated-choice polyhedral methods for continuous features might be improved with algorithms that seek non-zero choice balance.

Future Research

This paper, with its focus on stylized models, takes a different tack than recent algorithmic papers. The two tacks are complementary. Our stylized models can be extended to explain simulation results in Andrews, Ainslie and Currim (2002), empirical results in Moore (2003), or to explore the efficient use of self-explicated questions (Ter Hofstede, Kim and Wedel 2002). On the other hand, insights on utility balance, choice balance, and M-efficiency can be used to improve algorithms for both aggregate customization and polyhedral methods. Finally, interest in configurator-like questions is growing as more marketing research moves to the web and as e-commerce managers demand marketing research that matches the decisions they make daily. We have only begun to understand the nature of these questions.

References

- Andrews, Rick L., Andrew Ainslie, and Imran S. Currim (2002), "An Empirical Comparison of Logit Choice Models with Discrete versus Continuous Representations of Heterogeneity," *Journal of Marketing Research*, 39, (November), 479-487
- Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28, (September), 273-283.
- Dahan, Ely and John R. Hauser (2002), "The Virtual Customer," *Journal of Product Innovation Management*, 19, 5, (September), 332-354..
- Delquie, Philippe (2003), "Optimal Conflict in Preference Assessment," *Management Science*, 49, 1, (January), 102-115.
- Evgeniou, Theodoros, Constantinos Boussios, and Giorgos Zacharia (2003), "Generalized Robust Conjoint Estimation," Working Paper, (Fontainebleau, France: INSEAD), May.
- 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.
- Green, Paul E, Abba Krieger, and Manoj K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research*, pp. 215-222.
- _____, _____, and Yoram Wind (2001), "Thirty Years of Conjoint Analysis: Reflections and Prospects," *Interfaces*, 31, 3, Part 2, (May-June), S56-S73.
- Greene, William H. (1993), *Econometric Analysis*, 2E, (Englewood Cliffs, NJ: Prentice-Hall, Inc.)
- Haaiker, Rinus, Wagner Kamakura, and Michel Wedel (2000), "Response Latencies in the Analysis of Conjoint Choice Experiments," *Journal of Marketing Research* 37, (August), 376-382.
- Hauser, John R. and Steven M. Shugan (1980), "Intensity Measures of Consumer Preference," *Operations Research*, 28, 2, (March-April), 278-320.
- Huber, Joel and David Hansen (1986), "Testing the Impact of Dimensional Complexity and Affective Differences of Paired Concepts in Adaptive Conjoint Analysis," *Advances in Consumer Research*, 14, Melanie Wallendorf and Paul Anderson, eds., Provo, UT: Associations of Consumer Research, 159-163.
- _____ and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, 33, (August), 307-317.

- Johnson, Richard (1991), "Comment on `Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research*, 28, (May), 223-225.
- _____, Joel Huber, and Lynd Bacon, "Adaptive Choice Based Conjoint," *Sawtooth Software Proceedings*, 2003.
- Kanninen (2002), "Optimal Design for Multinomial Choice Experiments," *Journal of Marketing Research*, 39, (May), 214-227.
- Kuhfeld, Warren F. , Randall D. Tobias, and Mark Garratt (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, 31, 4, (November), 545-557.
- Lord, Frederic M. (1970), "Some Test Theory for Tailored Testing," Wayne H. Holzman (Ed.), *Computer Assisted Instruction, Testing, and Guidance*, (New York, NY: Harper and Row).
- Johnson, Richard (1991), "Comment on `Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research*, 28, (May), 223-225.
- _____, Joel Huber, and Lynd Bacon (2003), "Adaptive Choice Based Conjoint," *Proceedings of the Sawtooth Software Conference*, April.
- Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. C. Lee (1985), *The Theory and Practice of Econometrics*, (New York, NY: John Wiley and Sons).
- Keeney, Ralph and Howard Raiffa (1976), *Decisions with Multiple Consequences: Preferences and Value Tradeoffs*, (New York, NY: John Wiley & Sons).
- Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky (1971), *Foundations of Measurement*, (New York, NY: Academic Press).
- Liechty, John, Venkatram Ramaswamy, Steven Cohen (2001), "Choice-Menus for Mass Customization: An Experimental Approach for Analyzing Customer Demand With an Application to a Web-based Information Service," *Journal of Marketing Research*, 38, 2, (May).
- Louviere, Jordan J., David A. Hensher, and Joffre D. Swait (2000), *Stated Choice Methods: Analysis and Application*, (New York, NY: Cambridge University Press), pp. 354-381.
- Mahajan, Vijay and Jerry Wind (1992), "New Product Models: Practice, Shortcomings and Desired Improvements," *Journal of Product Innovation Management*, 9, 128-139.
- Marshall, Pablo, and Eric T. Bradlow (2002), "A Unified Approach to Conjoint Analysis Models," *Journal of the American Statistical Association*, 97, 459 (September), 674-682.
- McFadden, Daniel (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, P. Zarembka, ed., (New York: Academic Press), 105-142.

- Moore, William L. (2003), "A Cross-Validity Comparison of Conjoint Analysis and Choice Models," Working Paper, (Salt Lake City, Utah: University of Utah), February.
- Orme, Bryan (1999), "ACA, CBC, or Both: Effective Strategies for Conjoint Research," Working Paper, Sawtooth Software, Sequim, WA.
- _____ and Joel Huber (2000), "Improving the Value of Conjoint Simulations," *Marketing Research*, 12, 4, (Winter), 12-20.
- Sandor, Zsolt and Michel Wedel (2001), "Designing Conjoint Choice Experiments Using Managers' Prior Beliefs," *Journal of Marketing Research*, 38, 4, (November), 430-444.
- _____ and _____ (2002), "Profile Construction in Experimental Choice Designs for Mixed Logit Models," *Marketing Science*, 21, 4, (Fall), 455-475.
- _____ and _____ (2003), "Differentiated Bayesian Conjoint Choice Designs," working paper, University of Michigan Business School.
- Sawtooth Software (2002), "ACA 5.0 Technical Paper," Sawtooth Software Technical Paper Series, (Sequim, WA: Sawtooth Software, Inc.)
- Shugan, Steven M. (1980), "The Cost of Thinking," *Journal of Consumer Research*, 7, 2, (September), 99-111.
- Ter Hofstede, Frenkel, Youngchan Kim, and Michel Wedel (2002), "Bayesian Prediction in Hybrid Conjoint Analysis," *Journal of Marketing Research*, 39, (May), 253-261.
- Thomke, Stefan and Eric von Hippel (2002), "Customers as Innovators: A New Way to Create Value," *Harvard Business Review*, (April), 74-81.
- Toubia, Olivier, John R. Hauser, and Duncan I. Simester (2003), "Polyhedral Methods for Adaptive Choice-based Conjoint Analysis," forthcoming, *Journal of Marketing Research*.
- _____, Duncan Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," forthcoming, *Marketing Science*.
- Urban, Glen L. and John R. Hauser (2003), "'Listening In' to Find and Explore New Combinations of Customer Needs," forthcoming, *Journal of Marketing*.
- von Hippel, Eric (2001b), "Perspective: User Toolkits for Innovation," *Journal of Product Innovation Management*, 18, 247-257.

Appendix

Relative Bias Due to Metric Utility Balance

Proposition 1. For the stylized model, on average, metric utility balance biases partworths upward and does so differentially depending upon the true values of the partworths.

Proof. In the text we have already shown that one of the partworths is biased upwards for any zero-mean distribution that has non-zero measure below $p_2 - p_1$. We must now demonstrate that $\frac{\partial}{\partial p_1} E[e_{ub} | e_{ub} \geq p_2 - p_1] > 0$ for $p_1 - p_2 \geq 0$ with density functions which have non-zero density below $p_2 - p_1$. We differentiate the integral to obtain:

$$\begin{aligned} \frac{\partial}{\partial p_1} E[e_{ub} | e_{ub} \geq p_2 - p_1] &= \frac{\partial}{\partial p_1} \left[\frac{\int_{p_2-p_1}^{\infty} e_{ub} f(e_{ub}) de_{ub}}{\int_{p_2-p_1}^{\infty} f(e_{ub}) de_{ub}} \right] \\ &= \frac{\left(\int_{p_2-p_1}^{\infty} f(e_{ub}) de_{ub} \right) (p_2 - p_1) f(p_2 - p_1) - \left(\int_{p_2-p_1}^{\infty} e_{ub} f(e_{ub}) de_{ub} \right) f(p_2 - p_1)}{\left(\int_{p_2-p_1}^{\infty} f(e_{ub}) de_{ub} \right)^2} = \\ &= \frac{-[1 - F(p_2 - p_1)](p_1 - p_2) f(p_2 - p_1) - E[e_{ub} | e_{ub} \geq p_2 - p_1][1 - F(p_2 - p_1)] f(p_2 - p_1)}{[1 - F(p_2 - p_1)]^2} < 0 \end{aligned}$$

where the last step recognizes that both terms in the numerator are negative whenever $f(e_{ub})$ has density below $p_2 - p_1$ and $p_1 - p_2 > 0$. When $p_1 = p_2$, the second term is still negative.

Choice Balance vs. Magnitude and Non-focal Features

From Equation 3 in the text we know that $\text{trace}(\Sigma^{-1}) = \sum_{i=1}^q \sum_{k=1}^K d_{ik}^2 P_{i1} (1 - P_{i1})$. Because this function is separable in i , we focus on a single i and drop the i subscript. Rewriting P_i in terms of the partworths, \vec{p} , and allowing m to scale the magnitude of the

partworths, we obtain for a given i , $T \equiv \text{trace}_i(\Sigma^{-1}) = \sum_{k=1}^K d_k^2 / (e^{m\bar{d}\bar{p}} + e^{-m\bar{d}\bar{p}} + 2)$. We as-

sume, without loss of generality, that $p_K > 0$ and $y_{K-1} \equiv \sum_{k=1}^{K-1} p_k d_k \leq 0$. Define

$$s_{K-1} \equiv \sum_{k=1}^{K-1} d_k^2 \text{ and let } d_K^* \text{ be the optimal } d_K.$$

Lemma 1. For $p_K > 0$ and $y_{K-1} \leq 0$, $d_K^ \geq 0$.*

Proof. Assume $d_K^* < 0$. If $a^* = p_K d_K^* + y_{K-1} < 0$, then $\text{trace}^* = (d_K^{*2} + s_{K-1}) / (e^{ma^*} + e^{-ma^*} + 2)$. Consider $d_K^{**} = (-a^* - y_{K-1}) / p_K > 0$ such that $a^{**} = -a^* > 0$. This assures that the denominator of the trace stays the same. Now $|d_K^{**}| > |(-a^* - y_{K-1}) / p_K| > |(a^* - y_{K-1}) / p_K| = |d_K^*|$ because both a^* and y_{K-1} are of the same sign. Thus, the numerator of the trace is larger and the denominator is unchanged and we have the result by contraction. In the special case of $y_{K-1} = 0$, $\text{trace}(d_K) = \text{trace}(-d_K)$, hence we can also restrict ourselves to $d_K \geq 0$.

Lemma 2. For $p_K > 0$, $y_{K-1} < 0$, then the trace has no minimum in d_K on $(0, \infty)$.

Proof.

$T'(d_K) = [2d_K(e^{m\bar{d}\bar{p}} + e^{-m\bar{d}\bar{p}} + 2) - \sum_k d_k^2 m p_K (e^{m\bar{d}\bar{p}} - e^{-m\bar{d}\bar{p}})] / (e^{m\bar{d}\bar{p}} + e^{-m\bar{d}\bar{p}} + 2)^2 \equiv h(d_K) / b(d_K)$. Then, $T'(d_K^*) = 0 \Rightarrow h(d_K^*) = 0$ and $T''(d_K^*) = h'(d_K^*) / b(d_K^*)$. $T''(d_K^*)$ will have the same sign as

$$h'(d_K^*) = 2(e^{m\bar{d}\bar{p}} + e^{-m\bar{d}\bar{p}} + 2) + 2d_K m p_K (e^{m\bar{d}\bar{p}} - e^{-m\bar{d}\bar{p}}) - 2d_K m p_K (e^{m\bar{d}\bar{p}} - e^{-m\bar{d}\bar{p}}) - \sum_k d_k^2 m^2 p_K^2 (e^{m\bar{d}\bar{p}} + e^{-m\bar{d}\bar{p}}).$$

Cancel the second and third terms. Using the FOC gives

$$h'(d_K^*) = \sum_k d_k^2 m p_K (e^{m\bar{d}\bar{p}} - e^{-m\bar{d}\bar{p}}) / d_K - \sum_k d_k^2 m^2 p_K^2 (e^{m\bar{d}\bar{p}} + e^{-m\bar{d}\bar{p}}).$$

Thus, $h'(d_K^*)$ and $H \equiv (e^{m\bar{d}\bar{p}} - e^{-m\bar{d}\bar{p}}) - m d_K p_K (e^{m\bar{d}\bar{p}} + e^{-m\bar{d}\bar{p}})$ have the same sign. Because $d_K^* \geq 0$ by assumption, the FOC imply that $e^{m\bar{d}\bar{p}} - e^{-m\bar{d}\bar{p}} \geq 0$, hence $m\bar{d}\bar{p} > 0$. If $m d_K p_K \geq 1$, then H

sumption, the FOC imply that $e^{m\vec{d}\vec{p}} - e^{-m\vec{d}\vec{p}} \geq 0$, hence $m\vec{d}\vec{p} > 0$. If $md_K p_K \geq 1$, then $H < 0$ and so is $T''(d_K^*)$. If $md_K p_K < 1$, then $m\vec{d}\vec{p} < 1$ because $y_{K-1} < 0$. Thus, $e^0 < e^{m\vec{d}\vec{p}} < e^1$, hence $H < e - 1 - 2md_K p_K$. Thus, $H < 0$ if $md_K p_K > (e - 1)/2$. Call $h_0 = (e - 1)/2$. If $md_K p_K < h_0$, then $0 < m\vec{d}\vec{p} < h_0$ and $H < e^{h_0} - 1 - 2md_K p_K$, which is negative if $md_K p_K > h_1 > (e^{h_0} - 1)/2$. For all h in $(0, 1]$, we have $(e^h - 1)/2 > h$, so by recursion we show that $H < 0$ if $md_K p_K > h_\ell$ and h_ℓ converging to zero.

Proposition 2. As response accuracy increases, choice balance increases and the level difference in the focal feature decreases.

Proof. Assume $p_K > 0$ and $y_{K-1} \leq 0$ without loss of generality and rewrite $T = u(d_K, m)/v(d_K, m)$ where $u(d_K, m) = \sum_k d_k^2$, $v(d_K, m) = (e^{m\vec{d}\vec{p}} + e^{-m\vec{d}\vec{p}} + 2)$. For $m_o > 0$, $d_K^*(m_o)$ satisfies $f \equiv u'/u = v'/v \equiv g$. The numerator, u , does not depend upon m . Taking derivatives yields $v'/v = mp_K(e^{m\vec{d}\vec{p}} - e^{-m\vec{d}\vec{p}})/(e^{m\vec{d}\vec{p}} + e^{-m\vec{d}\vec{p}} + 2)$ and $\partial(v'/v)/\partial m =$

$$\frac{[p_K(e^{m\vec{d}\vec{p}} - e^{-m\vec{d}\vec{p}}) + mp_K\vec{d}\vec{p}(e^{m\vec{d}\vec{p}} + e^{-m\vec{d}\vec{p}})][e^{m\vec{d}\vec{p}} + e^{-m\vec{d}\vec{p}} + 2] - mp_K\vec{d}\vec{p}(e^{m\vec{d}\vec{p}} - e^{-m\vec{d}\vec{p}})^2}{(e^{m\vec{d}\vec{p}} + e^{-m\vec{d}\vec{p}} + 2)^2}.$$

We rearrange the numerator to obtain the following expression:

$p_K(e^{m\vec{d}\vec{p}} - e^{-m\vec{d}\vec{p}})(e^{m\vec{d}\vec{p}} + e^{-m\vec{d}\vec{p}} + 2) + 2mp_K\vec{d}\vec{p}(e^{m\vec{d}\vec{p}} + e^{-m\vec{d}\vec{p}}) + 4mp_K\vec{d}\vec{p}$, which is positive because $m\vec{d}\vec{p} > 0$ and $e^{m\vec{d}\vec{p}} - e^{-m\vec{d}\vec{p}} \geq 0$ as proven in Lemma 2. Hence, v'/v is increasing in m . Thus, $g(d_K^*(m_o), m_1) > g(d_K^*(m_o), m_o)$ for $m_1 > m_o$, which implies $f(d_K^*(m_o), m_1) = f(d_K^*(m_o), m_o) = g(d_K^*(m_o), m_o) < g(d_K^*(m_o), m_1)$. Hence, $T'(d_K^*(m_o), m_1) < 0$.

We have shown that at $m = m_1$, the derivative to T with respect to d_K is negative at $d_K^*(m_o)$. By Lemma 2, it is non-positive for all $d_K > d_K^*(m_o)$. Thus, $d_K^*(m_1) < d_K^*(m_o)$ which proves the second part of the proposition. Because

$y_{K-1} + d_K^*(m_o)p_K > 0$, $y_{K-1} + d_K^*(m_1)p_K > 0$, and $y_{K-1} < 0$, we have $0 < y_{K-1} + d_K^*(m_1)p_K < y_{K-1} + d_K^*(m_o)p_K$, which proves the first part of the proposition.

Proposition 3. As the levels of the non-focal features increase, choice balance increases.

Proof. We now consider changes in the non-focal features. In this case, m scales the non-focal features but does not scale d_K . Thus,

$T = (m^2 y_{K-1} + d_K^2) / (e^{my_{K-1} + d_K p_K} + e^{-(my_{K-1} + d_K p_K)} + 2)$. We again rewrite

$T = u(d_K, m) / v(d_K, m)$. Caution, the definitions of u and v (and f and g) are within the proofs and vary between the two propositions. Lemmas 1 and 2 still hold because the proofs follow the same principles. Thus, w.l.o.g., we have $p_K > 0$, $y_{K-1} \leq 0$ which implies that $d_K^* \geq 0$ and that $my_{K-1} + d_K^* p_K > 0$. This is true for any $m > 0$.

Let $m_o > 0$, then the FOC imply $f(d_K^*, m_o) \equiv v' / v = u' / u \equiv g(d_K^*, m_o)$. Let $m_1 \equiv m_o + dm_o$ and consider a $d_K = d_K^*(m_o) + dd_K$ such that utility is unchanged: $a^*(m_o) \equiv m_o y_{K-1} + d_K^*(m_o) p_K = (m_o + dm_o) y_{K-1} + (d_K^*(m_o) + dd_K) p_K$. Solving we obtain $dd_K = -dm_o y_{K-1} / p_K$. For reference call this Expression 1. Differentiating w.r.t. d_K gives $f(d_K^*, m_o) = p_K (e^{a^*} - e^{-a^*}) / (e^{a^*} + e^{-a^*} + 2)$. Because a^* is unchanged we have $f(d_K^*(m_o) + dd_K, m_o + dm_o) = f(d_K^*(m_o), m_o)$.

Differentiating u w.r.t. d_K gives: $g(d_K^*, m_o) = 2d_K^* / (m_o^2 s_{K-1} + d_K^{*2})$. Further differentiation gives $\partial g(d_K^*, m_o) / \partial d_K = (2m_o^2 s_{K-1} - 2d_K^{*2}) / (m_o^2 s_{K-1}^2 + d_K^{*2})^2$ and

$\partial g(d_K^*, m_o) / \partial m = -4d_K^* m s_{K-1} / (m_o^2 s_{K-1} + d_K^{*2})^2$. Consider $dm_o > 0$. Using Taylor's Theorem, ignoring higher order terms because dm_o and, hence, dd_K are differentials:

$g(d_K^* + dd_K, m_o + dm_o) \approx g(d_K^*, m_o) + (2m_o^2 s_{K-1} - 2d_K^{*2}) / (m_o^2 s_{K-1}^2 + d_K^{*2})^2 dd_K - 4d_K^* m_o s_{K-1} / (m_o^2 s_{K-1} + d_K^{*2})^2 dm_o$. Expression 1

implies $G_o \equiv g(d_K^* + dd_K, m_o + dm_o) - g(d_K^*, m_o)$ is of the same sign as

$G_1 \equiv -4d_K^*(m_o) m_o s_{K-1} - (2m_o^2 s_{K-1} - 2d_K^{*2}(m_o)) y_{K-1} / p_K$. Because $y_{K-1} < 0$,

$G_1 < -4d_K^*(m_o)m_o s_{K-1} - 2m_o^2 s_{K-1} y_{k-1} / p_K$. Further, G_1 is of the same sign as
 $G_2 = -2d_K^*(m_o)p_K - m_o y_{K-1} < -(m_o y_{K-1} + d_K^*(m_o)p_K) < 0$. Thus, for $dm_o > 0$ we have
 $f(d_K^*(m_o) + dd_K, m_o + dm_o) = f(d_K^*(m_o), m_o) = g(d_K^*(m_o), m_o)$
 $> g(d_K^*(m_o) + dd_K, m_o + dm_o)$. Thus, if we increase m_o while adjusting d_K to maintain
the same level of choice balance (constant a^*), $T' < 0$. Hence,
 $0 < d_K^*(m_o + dm_o) < d_K^*(m_o) + dd_K$ and $0 < a^*(m_o + dm_o) < a^*(m_o)$. Thus, choice balance
increases.

Relative Efficiency of Configurator Questions for E-commerce Decisions

Lemma 3. If e_k , e_ℓ , and e_t are zero-mean, i.i.d. normally distributed, then
 $\text{Prob}[e_k \geq -p_k] > \text{Prob}[e_\ell \pm e_t \geq -p_k]$ for $k = 1, 2$.

Proof. If the errors are independent and identically distributed zero-mean normal
random variables with variances, σ^2 , then, for $p_k \geq 0$, the comparison reduces to the fol-
lowing equation where $f(\bullet|\sigma)$ is a normal density with variance σ^2 . Let $e_T = e_\ell \pm e_t$.

$$\frac{1}{2} + \int_{-p_k}^0 f(e_k | \sigma) de \geq \frac{1}{2} + \int_{-p_k}^0 f(e_T | \sigma\sqrt{2}) de_T \Leftrightarrow \int_0^c f(e_k | \sigma) de \geq \int_0^c f(e_T | \sigma\sqrt{2}) de \text{ with } c \geq 0$$

To demonstrate that the last expression is true, we let $I = \int_0^c f(e_T | \sigma\sqrt{2}) de =$
 $\int \exp(-e^2 / 4\sigma^2) / (\sigma\sqrt{4\pi}) de$. Change the variables in the integration setting $g = e / \sqrt{2}$.

Then,

$$I = \int_0^{c/\sqrt{2}} \exp(-g^2 / 2\sigma^2) / (\sigma\sqrt{2\pi}) dg \leq \int_0^c \exp(-e^2 / 2\sigma^2) / (\sigma\sqrt{2\pi}) de = \int_0^c f(e_T | \sigma) de.$$

The conditions for $p_k \leq 0$ yield equivalent conditions.

Proposition 4. For zero-mean normally distributed (i.i.d.) errors, the warm-
up questions (Q1+Q2) provide more accurate estimates of feature choice than do
pairs of questions that include a metric paired-comparison question (Q3).

Proof. With Q1+Q2 or Q1+Q3, the error in estimating \hat{p}_1 is based directly on e_1 and the probability of correct prediction is $\text{Prob}[e_1 \geq -p_1]$. With Q2+Q3 we obtain \hat{p}_1 by adding Q2 and Q3 and the error in estimating \hat{p}_1 is based on $e_2 + e_t$. The probability of correct prediction is $\text{Prob}[e_2 + e_t \geq -p_1]$. Thus, for Feature 1, Q1+Q2 and Q1+Q3 are superior to Q2+Q3 if $\text{Prob}[e_1 \geq -p_1] > \text{Prob}[e_2 + e_t \geq -p_1]$. For Feature 2, similar arguments show that Q1+Q2 and Q2+Q3 are superior to Q1+Q3 if $\text{Prob}[e_2 \geq -p_2] > \text{Prob}[e_1 - e_t \geq -p_2]$. Thus, Q1+Q2 is superior to Q2+Q3 and Q1+Q3 if the condition in the text holds. Finally, by Lemma 3 this condition holds for zero-mean, i.i.d., normal errors.