An Optimization Framework for the Adaptive Design of Robust Choice Questionnaires

JACOB ABERNETHY, THEODOROS EVGENIOU, OLIVIER TOUBIA, and JEAN-PHILIPPE VERT [1]

[1] Jacob Abernethy is a graduate student at Toyota Technological Institute at Chicago, Chicago, IL 60637, USA (e-mail: jabernethy@tti-c.org) and was a researcher at INSEAD during this work, Theodoros Evgeniou is an Assistant Professor of Technology Management at INSEAD, Bd de Constance, Fontainebleau 77300, France (Tel: +33 (0)1 60 72 45 46, Fax: +33 (0)1 60 74 55 01, e-mail: theodoros.evgeniou@insead.edu); Olivier Toubia is an Assistant Professor at the Marketing Division at Columbia Business School, 3022 Broadway, Room 522, New York, NY 10027, USA (e-mail: ot2107@columbia.edu); and Jean-Philippe Vert is a Researcher at Ecole des Mines de Paris, 35 rue Saint-Honoré, Fontainebleau 77300, France (e-mail: jean-philippe.vert@mines.org).

An Optimization Framework for the Adaptive Design of Robust Choice Questionnaires

Abstract

We propose a general framework for adaptively designing choice-based conjoint questionnaires at the individual level. This framework uses complexity control to improve the robustness of the conjoint designs to response error and links the informativeness of conjoint questions to the Hessian of the loss function minimized in partworth estimation. It formalizes and generalizes several methods recently proposed both for questionnaire design and estimation. Simulations as well as an online experiment suggest that it outperforms established benchmarks, especially when response error is high.

Keywords: Choice Models, Marketing Research, Data Mining, Regression And Other Statistical Techniques, Marketing Tools.

1 Introduction

An important problem in marketing, and in particular in conjoint analysis, is the design of questionnaires that can effectively capture the preferences of individuals. This can be approached by designing the conjoint questions adaptively, based on previous information. Aggregate customization, for example, adapts the questions *across* respondents (Arora and Huber 2001; Huber and Zwerina 1996; Sandor and Wedel 2001). Other methods perform the adaptation *within* respondents. The development of web-based questionnaires has recently increased the interest in such adaptive methods, among practitioners as well as academics (Hauser, Tellis and Griffin, 2005). Two notable examples are Adaptive Conjoint Analysis (Johnson 1987; Sawtooth Software 1996, 2002) and polyhedral methods (Toubia et al., 2003; Toubia et al., 2004). Note that only one of these within-respondent methods (Toubia et al., 2004) deals with the increasingly popular format of choice based conjoint analysis.

As noted by Hauser and Toubia (2005), adaptive designs are by definition subject to endogeneity, because the questions are influenced by the noise to the previous answers. As a result, adaptive methods have been shown to be sensitive to response error. In particular, while previously proposed adaptive methods tend to outperform non-adaptive benchmarks when response error is low, they typically do not perform as well when response error is high (see for example the simulations of Toubia et al., 2003 and Toubia et al., 2004). This suggests the need for designs that retain the benefits from adaptive interviews while being robust to response error (i.e., less affected by endogeneity).

In this paper we propose a general framework for constructing such designs. Robustness to response error is achieved by extending complexity control, used so far only for *estimation*, to the *design* of questionnaires. Another contribution of this paper is to propose a link between the informativeness of conjoint questions and the Hessian of the loss function minimized by the estimation procedure. This link also formalizes and generalizes previously proposed heuristics such as the ones used in polyhedral methods (Toubia et al., 2004).

Our approach is motivated by the well established field of statistical learning theory (Vapnik 1998), much like recently proposed conjoint *estimation* methods (Cui and Curry 2005; Evgeniou et al., 2005) and not unlike previous optimization approaches to conjoint estimation (Srinivasan and Shocker 1973; Srinivasan 1998) (those methods were developed *only* for the estimation of preference models from existing data and *not* for the design of questionnaires).

The paper is organized as follows. We first present the framework in Section 2. We then test it using simulations (described in Section 3) and an online experiment (reported in Section 4). Section 5 concludes and suggests areas for future research.

2 Adaptive Design of Robust Choice Questionnaires

Any conjoint analysis method can be viewed as being composed of two key building blocks:

1. A method for designing questionnaires;

2. A method for estimating the respondents' partworths based on their responses to the questionnaires.

For ease of exposition we will first discuss (2) and then introduce our general framework for questionnaire design, which will be our main focus.

2.1 Notations

We make the standard assumption (Ben-Akiva and Lerman 1985; Srinivasan and Shocker 1973) of additive utility functions.[2] We denote a profile with a *row* vector $\mathbf{x}$ and an individual's partworths with a *column* vector $\mathbf{w}$, such that his or her utility for a profile denoted by $\mathbf{x}$ is $U(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w}$. For simplicity we first consider binary choices, such that question i consists of two alternatives $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}\}$. We will later discuss the generalization to non-binary choices. Our goal is to adaptively design the $(n+1)^{th}$ question (pair of profiles) for an individual using the information contained in his or her first n choices. Without loss of generality, we assume that in each question i the respondent chooses the first product $\mathbf{x}_{i1}$.

2.2 Robust Estimation

A standard and traditional approach to estimation is to maximize some measure of fit. However this may lead to overfitting and the estimates may be sensitive to noise, especially

[2] For simplicity we will not address interactions here, although an important advantage of our approach is that it allows an easy and computationally efficient estimation of interactions, using the kernel transformations introduced in the Support Vector Machines literature (Cortes and Vapnik 1995; Vapnik 1998; Cui and Curry 2005; Evgeniou et al., 2005).

if they are based on limited data (e.g., few choices per respondent). Statistical learning theory (Vapnik 1998) has addressed this issue by introducing the fundamental notion that the estimates should reflect a trade off between maximizing fit and minimizing complexity. Complexity is typically defined as deviation from a null model. The term "complexity control" comes from the fact that this trade off essentially limits the set of possible estimates, making this set less "complex" (e.g., smaller). It has been established (see for example Cui and Curry 2005; Evgeniou et al., 2005; Vapnik 1998) that complexity control yields *estimates* that are more robust to noise. We extend the use of complexity control to the adaptive *design* of choice experiments and show that similar robustness results hold for *question selection*.

Before introducing our design method, let us first describe our estimation paradigm in more details. Given the answers to n choice questions $\{(\mathbf{x}_{11}, \mathbf{x}_{12}), (\mathbf{x}_{21}, \mathbf{x}_{22}), \ldots, (\mathbf{x}_{n1}, \mathbf{x}_{n2})\}$, where we assume that the first alternative $\mathbf{x}_{i1}$ was always preferred, statistical learning theory estimates the partworths $\mathbf{w}$ as the minimizer of a loss function of the following general form:

$$\mathbf{min}_{\mathbf{w}} \; V(\mathbf{w}, \{(\mathbf{x}_{i1}, \mathbf{x}_{i2})\}) + \lambda\Phi(\mathbf{w}), \tag{1}$$

where $V(\mathbf{w}, \{(\mathbf{x}_{i1}, \mathbf{x}_{i2})\})$ measures the fit and Φ controls (penalizes) the complexity of the partworths. Different specifications of V and Φ lead to different special cases of statistical learning theory methods. In this paper we adopt a specific formulation known as Regularization Networks (RN) (Tikhonov and Arsenin, 1977; Vapnik, 1998; Evgeniou et al., 2000). RN has the property that the corresponding loss function is convex and twice differentiable, which will be essential to our questionnaire design approach. Moreover, it leads to closed form solutions that are fast to compute.[3] RN estimation solves the following minimization problem:

$$\mathbf{min_w}\ R_n(\mathbf{w}) = \sum_{i=1\ldots n} (1 - (\mathbf{x}_{i1} - \mathbf{x}_{i2}) \cdot \mathbf{w})^2 + \lambda \|\mathbf{w}\|^2 \tag{2}$$

With this formulation, fit between the estimated utilities and the observed choices is measured by $\sum_{i=1\ldots n}[1 - (\mathbf{x}_{i1} - \mathbf{x}_{i2}) \cdot \mathbf{w}]^2$. The constant of 1 plays a scaling role for $\mathbf{w}$. More precisely, with this definition fit is maximized if the choices are satisfied with a margin as close as possible to 1. The second part of Equation (2), $\|\mathbf{w}\|^2$, controls the complexity of the partworth solution $\mathbf{w}$. This formulation of complexity, which is by far the most common in statistical learning theory (Vapnik 1998), was also adopted by Cui and Curry (2005) and Evgeniou et al (2005). With this formulation complexity can be viewed as deviation from a null model in which all the partworths are equal ($\mathbf{w} = 0$).[4] The parameter λ reflects the trade-off between fit and complexity, and is typically set by the researcher (we set it to $\frac{1}{n}$ – see below). This parameter may also be chosen using cross-validation or a validation set (Evgeniou et al., 2005; Vapnik 1998).

[3] Cui and Curry (2005) and Evgeniou et al. (2005) used another special case of (1) known as Support Vector Machines (SVM) (Cortes and Vapnik 1995; Vapnik 1998). Our framework does not apply to SVM because that loss function is not twice differentiable. However, we note that RN estimation has been shown to perform similarly to SVM estimation in many applications (e.g. Rifkin, 2004).

[4] Note that complexity here does not carry any cognitive meaning. In Section 5 we propose that future research may focus on the "cognitive complexity" of the choice process.

2.3 Robust Adaptive Questionnaire Design

We now extend the use of complexity control to the adaptive design of robust choice-based conjoint questionnaires. Our approach is based on the analysis of the effect of a new question on the estimate of the partworths.

2.3.1 Intuition

Let us assume that n questions have been asked thus far, and that the loss function R_n given by Equation (2) is minimized by our estimate $\mathbf{w}_n$.[5] Let us denote by $\mathbf{z}_i = (\mathbf{x}_{i1} - \mathbf{x}_{i2})$ the difference (row) vector between the two profiles in question i. Notice that the loss function R_n only depends on the $\mathbf{z}_i$'s. Furthermore, if the product attributes were real-valued, we would only need to determine $\mathbf{z}_{n+1}$ in order to generate the next question (and choose any two profiles such that $\mathbf{z}_{n+1} = (\mathbf{x}_{n+1,1} - \mathbf{x}_{n+1,2})$). We first develop the intuition for real-valued attributes and later discuss how to design profiles with discrete attribute levels. In creating the next question, we consider two criteria:

1. *Choose a direction for $\mathbf{z}_{n+1}$ along which the current loss function $R_n(\mathbf{w})$ is as flat as possible.* The flatness of the loss function is by definition given by its second derivate matrix, the Hessian (see details in the next section). The rationale behind our criterion is that the flatness of the loss function may be interpreted as a measure of uncertainty in the partworth estimates. For example, in the case of maximum likelihood estimation, the Hessian of the loss function is asymptotically equal to the

[5] Although we focus on the RN formulation in this paper, our approach can be applied to any loss function (1) that is convex and twice differentiable with respect to $\mathbf{w}$.

inverse covariance matrix of the estimates (Newey and McFadden 1994).[6] Asking a new question along the flattest direction of the loss function will have the highest impact on the estimated utility function and will yield the greatest reduction in partworth uncertainty. In that sense it will be most informative.

2. *Utility balance.* We attempt to create a question involving a set of products that are as equally attractive as possible. In the case of binary choices, this implies: $\mathbf{x}_{n+1,1} \cdot \mathbf{w}_n \approx \mathbf{x}_{n+1,2} \cdot \mathbf{w}_n$, or more simply, $\mathbf{z}_{n+1} \cdot \mathbf{w}_n \approx 0$. Utility Balance is a common criterion in the literature, and has been shown to increase the statistical efficiency of choice designs (Arora and Huber 2001; Huber and Zwerina 1996; Toubia et al., 2004; Kanninen 2002; Hauser and Toubia 2005).

2.3.2 Hessian-based question selection

Let us now formalize our two criteria. First, let us assume, as is the case with the RN formulation of Equation (2), that $R_n(\mathbf{w})$ is strictly convex and twice differentiable. Formally, the estimated utility vector is the only point $\mathbf{w}_n$ that satisfies:

$$\nabla R_n\left(\mathbf{w}_n\right) = 0.$$

[6] Although it is beyond the scope of this paper, a link can be made between RN and maximum likelihood estimation. See Evgeniou et al. (2000) for details.

Where ∇ is the gradient operator. Around that minimum, the "flatness" (or convexity) of R_n is given by its second derivative matrix (Hessian):

$$\left[\nabla^2 R_n\right]_{i,j} := \frac{\partial^2 R_n}{\partial w_i \partial w_j}.$$

More precisely, the convexity along a direction $\mathbf{z}$ is given by $\mathbf{z}\nabla^2 R_n \mathbf{z}^\top$.

In order to find the direction of smallest convexity (criterion 1) orthogonal to $\mathbf{w}_n$ (criterion 2), we therefore solve the following optimization problem:

$$\mathbf{min_z} \quad \mathbf{z}\nabla^2 R_n\left(\mathbf{w}_n\right)\mathbf{z}^\top \tag{3}$$

Subject to

$$\mathbf{z}\mathbf{w}_n = 0,$$

$$\mathbf{z}\mathbf{z}^\top = 1,$$

where $\mathbf{z}\mathbf{z}^\top = 1$ is a scaling constraint. After projecting the Hessian matrix onto the hyperplane orthogonal to $\mathbf{w}_n$ by the equation:

$$B_n := \left(\mathbf{I}_p - \frac{\mathbf{w}_n \mathbf{w}_n^\top}{\mathbf{w}_n^\top \mathbf{w}_n}\right)\nabla^2 R_n\left(\mathbf{w}_n\right), \tag{4}$$

where p is the dimensionality of $\mathbf{w}_n$ and $\mathbf{I}_p$ is the $p \times p$ identity matrix, this problem reduces to finding the eigenvector $\hat{\mathbf{z}}_{n+1}$ associated with the smallest positive eigenvalue of B_n.

Thus stated, this strategy is very general and can be applied to any estimation procedure of the form (1) as long as V and Φ are such that the loss function is convex and twice differentiable. For Regularization networks (RN) defined in Equation (2), it can be shown (see Appendix A) that the estimate $\mathbf{w}_n$ after n questions is

$$\mathbf{w}_n = \left(Z_n^\top Z_n + \lambda \mathbf{I}_p\right)^{-1} Z_n^\top \mathbf{1}_n, \tag{5}$$

where $\mathbf{1}_n$ is a vector of 1's and Z_n is the design matrix after n questions (the i^{th} row of Z_n is $Z_{ni} = (\mathbf{x}_{i1} - \mathbf{x}_{i2})$). The next question is designed using the eigenvector associated with the smallest positive eigenvalue of:

$$\left(\mathbf{I}_p - \frac{\mathbf{w}_n \mathbf{w}_n^\top}{\mathbf{w}_n^\top \mathbf{w}_n}\right)\left(Z_n^\top Z_n + \lambda \mathbf{I}_p\right), \tag{6}$$

In summary, when coupled with RN estimation, the proposed conjoint analysis method consists of the following two steps at each iteration n:

1. *Step 1:* Compute the estimate of the partworths given by Equation (5)

2. *Step 2:* The next question (difference vector $\mathbf{z}_{n+1}$) is defined by the eigenvector associated with the smallest positive eigenvalue of the matrix given by Equation (6).

Note that all the expressions are in closed form and only require the inversion of a matrix of size equal to the number of partworths. Hence this method is very fast computationally.

2.4 Practical Issues

Before proceeding to the validation of the method, we briefly discuss some practical implementation issues:

- **Designing the first question:** Before the first question, most of the positive eigenvalues of the Hessian are equal, i.e., there are many "smallest positive eigenvalues". As in previous work (e.g., Toubia et al., 2004) we design the first question randomly.

- **Designing questions with more than 2 profiles:** When more than two profiles per question are needed we consider not only the smallest positive eigenvalue of the Hessian (4) but also the second smallest, third, etc. We illustrate in Appendix B the case of 4 profiles per question.

- **Choosing the parameter λ in (1):** As discussed above (see also Vapnik 1998), the "trade off" parameter λ in (1) is often chosen in practice using a small validation set or using cross-validation. While this is feasible *ex post* when estimating the partworths, this is not feasible *ex ante* when designing questions. In this paper we set λ to $\frac{1}{n}$, where n is the number of questions. This formulation addresses the concern that λ should decrease as the amount of data increases (Vapnik 1998). We leave other methods for determining λ (e.g., using data from another group of individuals) to future research.

- **Designing profiles with finite attribute levels:** The approach outlined above generates a continuous difference vector $\mathbf{z}_{n+1}$. In most cases, however, attribute levels are discrete and it is not possible to find two profiles such that $\mathbf{z}_{n+1} = (\mathbf{x}_{n+1,1} - \mathbf{x}_{n+1,2})$. We address this issue using the Knapsack problem proposed by Toubia et al. (2004). See Appendix B for more details.

- **Adding extra constraints:** Additional information about the utility vectors can be captured using virtual examples as discussed by Evgeniou et al. (2005). We note that there exists a large literature on the use of constraints in statistical learning theory methods (see for example Scholkopf et al., 1996).

2.5 Polyhedral estimation as a limit case

We show in Appendix C that the polyhedral *estimation* method of Toubia et al. (2004) can actually be written as a limit case of Equation (1) in which the weight on fit goes to $+\infty$. With an infinite weight on fit, estimation becomes a 2-stage procedure in which fit is maximized in the first step and complexity is minimized in the second step. In other words, maximizing fit is imposed as a constraint and the final estimate is the least complex vector among those that maximize fit. Therefore, a proper trade off between fit and complexity control is not achieved, leading to higher sensitivity to response error.

Toubia et al. (2004)'s *question selection* method also selects a direction of largest uncertainty (see Appendix C for a detailed comparison of the two methods). However, in their case the sensitivity of analytic center estimation to response error also carries over to polyhedral *question selection* (as will be confirmed by our experiments).

3 Simulations

3.1 Simulation Design

We first tested our approach using Monte Carlo simulations (Carmone and Jain, 1978; Andrews et al., 2002). We compared the performance of the following 4 types of conjoint designs (2 adaptive and 2 non-adaptive) under different levels of noise and heterogeneity:

- An orthogonal design
- An aggregate customized design (Arora and Huber 2001; Huber and Zwerina 1996; Sandor and Wedel 2001)
- An adaptive questionnaire designed using the polyhedral method (POLY) of Toubia et al. (2004)
- An adaptive RN-based questionnaire designed using the method proposed in this paper.

We used the increasingly standard simulation setup introduced by Arora and Huber (2001) and used among others by Toubia et al. (2003), Toubia et al. (2004), and Evgeniou et al. (2005). Our $2 \times 2 \times 4 \times 3$ design allowed for two levels of response accuracy and two levels of respondent heterogeneity. In each response accuracy × heterogeneity subdesign, each question selection method listed above was estimated using 3 different estimation methods:

- The Analytic Center estimation (AC) method of Toubia et al. (2004)
- RN estimation (see Equation (2))
- Hierarchical Bayes estimation (HB)

In order to ensure complete orthogonal and aggregate customization designs, we followed Arora and Huber (2001) and used 16 choice questions, each containing 4 alternatives defined by 4 features with 4 levels each. In each response accuracy × heterogeneity subdesign, we generated 5 sets of 100 respondents, with partworths drawn from a normal distribution with mean $(-\beta, -\frac{1}{3}\beta, \frac{1}{3}\beta, \beta)$ for each attribute, and with variance σ_β^2. The parameter β is a magnitude parameter that controls response accuracy, and was set to 0.5 and 2 in the low-accuracy and high-accuracy cells respectively. The parameter σ_β^2 controls heterogeneity and was set respectively to $\sigma_\beta^2 = 0.5\beta$ and $\sigma_\beta^2 = 2\beta$ in the low and high heterogeneity cells.[7] Each choice among 4 profiles was made according to the logistic probabilities. Our performance metric was the Root Mean Square Error (RMSE) of the estimated partworths. Both estimated and true partworths were normalized before computing the RMSE such that the partworths of each attribute summed to 0 and that their absolute values summed to 1 (Toubia et al., 2004).

Note that the polyhedral method uses information about the lowest level for each attribute both in question design and estimation. We used this information as well for

[7] Those values were used in the published simulations mentioned above.

the RN estimation and question selection methods, by using virtual examples (Evgeniou et al., 2005). Evgeniou et al. (2005) also demonstrate that, in simulations, HB performs better if, for each respondent, we constrain the HB estimates so that the partworth of the lowest level of each feature is also the smallest partworth for that feature. Following them, we re-draw the partworths from the Metropolis Hastings algorithm until they satisfy these constraints. As in Arora and Huber (2001), aggregate customization was based on the true mean of the population distribution. Relabeling and swapping were used to improve utility balance.

3.2 Simulation Results

We summarize all the results in Table 1, where each row corresponds to a different questionnaire design method and each column to a different estimation method. We compare estimation methods first and question selection methods second. This distinction enables us to distinguish the effect of complexity control on question selection (new to this paper) from its effect on estimation (studied in previous research).

3.2.1 Estimation

Our results confirm previous findings (Evgeniou et al., 2005; Toubia et al., 2004) that hierarchical Bayes performs very well in simulations in which its assumptions are satisfied (logistic choice probabilities, normally distributed partworths). In our case, HB provides the (significantly) lowest RMSE in all 4 Magnitude × Heterogeneity cells. In turn, RN performs better than the other individual level estimation method, AC. Out of the 16 Magnitude × Heterogeneity × Question selection method cells, RN is significantly better (at the $p < 0.05$ level) in 11, tied in 2 and significantly worse in 3.[8]

3.2.2 Question Design

The more relevant comparisons for this paper are the comparisons between question selection methods. The results suggest that the RN-based method is the best overall: it is significantly best or tied with best in 8 of 12 Magnitude × Heterogeneity × Estimation method cells, and best or tied with best in 3 of the 4 Magnitude × Heterogeneity cells.

We have argued that one of the main contributions of our approach is to produce adaptive conjoint designs that are robust to response error. Two types of comparisons are possible in order to test this claim. Comparing RN to the other *adaptive* method (POLY) allows us to evaluate whether RN designs are more robust to noise than other adaptive designs that do not control for complexity. Comparing RN to the *non-adaptive* methods also allows us to evaluate conditions under which the benefits from robust adaptive questionnaires overweight endogeneity issues.

Let us start with the first comparison. RN is significantly better than POLY in 10, tied in 1 and significantly worse in 1 of the 12 Magnitude × Heterogeneity × Estimation method cells. More importantly, RN is significantly better than POLY in all 6 low magnitude (high response error) cells, irrespective of whether the *estimation* method uses complexity control. For example, using HB estimation (the best estimation method), RN

[8] Note that these significance tests are *not* reported in Table 1.

performs on average 11.9% better than POLY in the high-response-error conditions, while it performs only 2.9% better on average in the low-response-error conditions. Using the vocabulary of Toubia et al. (2004), once a wrong answer is given POLY will be searching for an estimate in the wrong half-space, i.e., the "true" utility vector will be "left" on the other side of the hyperplane defined by the wrongly answered question. This phenomenon is an illustration of the dependence of adaptive questions on the noise to previous answers (i.e., endogeneity) which characterizes all adaptive methods (Hauser and Toubia 2005). Our simulations confirm that in the case of POLY, this phenomenon is exacerbated when response error is higher. On the other hand, the presence of a complexity control in the RN loss function attenuates this dependence, giving rise to high performing designs even if response error is high. To draw an analogy with POLY, with RN-based adaptive questions the choice of a solution space in which to search for an estimate is not only driven by the answers to the previous questions, but also by the complexity of the space.

We finally compare RN to the *non-adaptive* benchmarks (orthogonal and customized designs). In the high magnitude conditions, RN (as well as POLY - hence *both adaptive* designs) systematically outperforms *both non-adaptive* benchmarks, confirming the attractiveness of adaptive methods when response error is low, established in previous simulations (e.g., Toubia et al., 2004). The more interesting comparisons are when response error is high (low magnitude). RN-based questionnaires perform better than *both non-adaptive* methods in 3 of the 6 low magnitude × Heterogeneity × Estimation method cells and 1 of 2 low magnitude × Heterogeneity cells. This suggests that with *robust* adaptive designs, the benefits from adaptive questionnaires can overweight endogeneity issues even when response error is high.

– Insert Table 1 about here –

4 An Online Experiment

4.1 Experimental Design

We next tested the proposed framework using an online experiment. The sets of question selection methods and estimation methods tested in this experiment were the same as those tested in the simulations (question selection methods: Orthogonal design, Aggregate customization,[9] POLY, and RN; estimation methods: AC, RN and HB). 500 respondents from an online panel were randomly assigned to one of the four question-selection-method conditions, resulting in 125 respondents per condition. Each respondent completed a 16 question design obtained by the corresponding method, followed by 4 randomly designed holdouts (the transition from the questionnaire to the holdouts was seamless), a filler task (a questionnaire on the compromise effect), and an external validity ranking task. In this last task the respondents were asked to rank 6 profiles, randomly selected from a 16 profile orthogonal design (different from the one used for the questionnaire). See Figure 1 for example screenshots from the experiment. The product

[9] The prior used by aggregate customization was obtained from a pretest involving 50 students from a large west coast university.

chosen for this experiment was digital cameras. We focused on 5 features with 4 levels each: Price ($500, $400, $300, $200), Resolution (2, 3, 4, or 5 Megapixels), Battery Life (150, 300, 450, or 600 pictures), Optical Zoom (2x, 3x, 4x, 5x), and Camera Size (SLR, Medium, Pocket, Ultra compact). The features were introduced and described to the respondents before the questionnaire.[10] Each choice question comprised 4 alternatives.

Like in the simulations, we estimated each design with each estimation method, and compare estimation methods first and question design methods second.

– Insert Figure 1 about here –

4.2 Experimental Results

We measured performance using the following 3 metrics:

1. The average number of holdouts (out of 4) correctly predicted ("Holdout hit rate")
2. The proportion of first choices correctly predicted in the external validity task ("Choice hit rate")
3. The correlation between predicted and observed rankings in the external validity task ("Choice correlation")

[10] It was assumed that the first level of each attribute was the least preferred, hence there was no need for self-explicated questions.

4.2.1 Estimation

The experiment confirmed the superiority of HB as an estimation method. See Table 3 in Appendix D for a summary of the results when the estimation is done using all 16 questions. The average HB performance across question selection methods is significantly higher than the average AC and RN performances, for all three performance metrics (p-values < 0.05).

4.2.2 Question Design

Given our sample size, considering only the estimates after 16 questions does not yield enough observations to discriminate statistically between question selection methods. An alternative approach, yielding a higher statistical power, was introduced by Toubia et al. (2003). It consists in pooling the performance measures calculated after intermediate questions (we used questions 3 to 16) in order to increase the total number of observations. For example, while the orthogonal design performs (non-significantly) best on all three metrics after 16 questions, it performs best only in 10 of the 14×3=42 resulting question number × performance metric combinations. In order to capture *persistent* trends in performance differences, we followed Toubia et al. (2003) and estimated the following model for choice correlation:[11]

[11] Unlike Toubia et al. (2003), we did not include the performance of an equal weights model in our specification, because 15 out of the 16 profiles used in the external validity task achieved the same utility based on such a model. We tried other controls such as the total time taken to complete the questionnaire, or the correlation between the final ranking and 1,2,3,4,5,6. The parameter estimates corresponding to those controls were in the expected direction (longer time lead to higher performance, and higher correlation with 1,2,3,4,5,6 to lower performance), and the controls did not change the comparisons across question selection methods.

$$Correlation_{rq} = \sum_{q=3}^{16} \alpha_q Question_q + \sum_{m=1}^{3} \beta_m QMethod_m + \epsilon_{rq} \qquad (7)$$

where r indexes the respondents, q indexes the questions, and *Question* and *QMethod* refer to dummy variables capturing the question number and the question design method effects. Given the consistent superiority of HB, Equation (7) was based on the HB estimates.

We estimated Equation (7) using OLS. We used similar model specifications for the holdout hit rates and the choice hit rates, estimated using ordinal regression and logistic regression respectively. The comparisons across question selection methods are summarized in Table 2, where ">" denotes a significant difference at the $p < 0.05$ level, and $\approx$ denotes no significant difference.

Table 2 indicates that RN is the only question selection method that is best or non-significantly different from best under all 3 performance metrics. It is interesting to note that the comparisons of question selection methods depend upon the performance metric used. One possible explanation may be that respondents adapt their choice heuristics and preference structures as a function of the validation task (e.g., holdout versus external validity or choice versus ranking), and that these different underlying choice processes have varying levels of congruency with the linear partworths obtained by the different question selection methods (Payne, Bettman, and Johnson 1988, 1993; Bettman, Luce and Payne 1998). We leave a deeper understanding of the influence of the validation task on conjoint comparisons to future research. Such research may also investigate to what extent the performance of the RN-based method is driven by the fact that its designs are more robust not only to response error, but also to noise coming from misspecifications of the assumed linear model.

– Insert Table 2 about here –

5 Conclusions

We have proposed a general framework for designing *robust adaptive* choice-based conjoint questionnaires. We showed that complexity control, previously used only for estimation, can also be used to produce designs that are less affected by response error and endogeneity, which can then be estimated using any established method such as HB. Another contribution of this paper is to propose a link between the informativeness of choice questions and the Hessian of the loss function minimized in estimation. This link formalizes and generalizes previous heuristics for adaptive questionnaire design.

Both simulations and an online experiment suggest that the proposed method performs well compared to established benchmarks, and confirm that its main advantage lies in its integrated treatment of response error.

Various research questions can be explored in the future. On the more technical side, one could explore ways of better tuning the parameter λ adaptively as respondents answer questions. Another exciting area for future research is the generalization of the framework to non-compensatory processes (Allenby and Gilbride 2004; Yee et al., 2005; Kohli and

Jedidi 2005). The loss function in Equation (1) could be generalized to include the ”cognitive” complexity of the choice process. Indeed, a fundamental premise of the work of Vapnik (1998), Cui and Curry (2005), and Evgeniou et al. (2005) is that constraints (e.g., in the forms of penalties) on the partworths (such as the complexity control $\|w\|^2$) lead to estimates that are more accurate and robust to noise. Further constraints (e.g., other complexity controls), based on prior knowledge about how people make choices, may further improve performance.

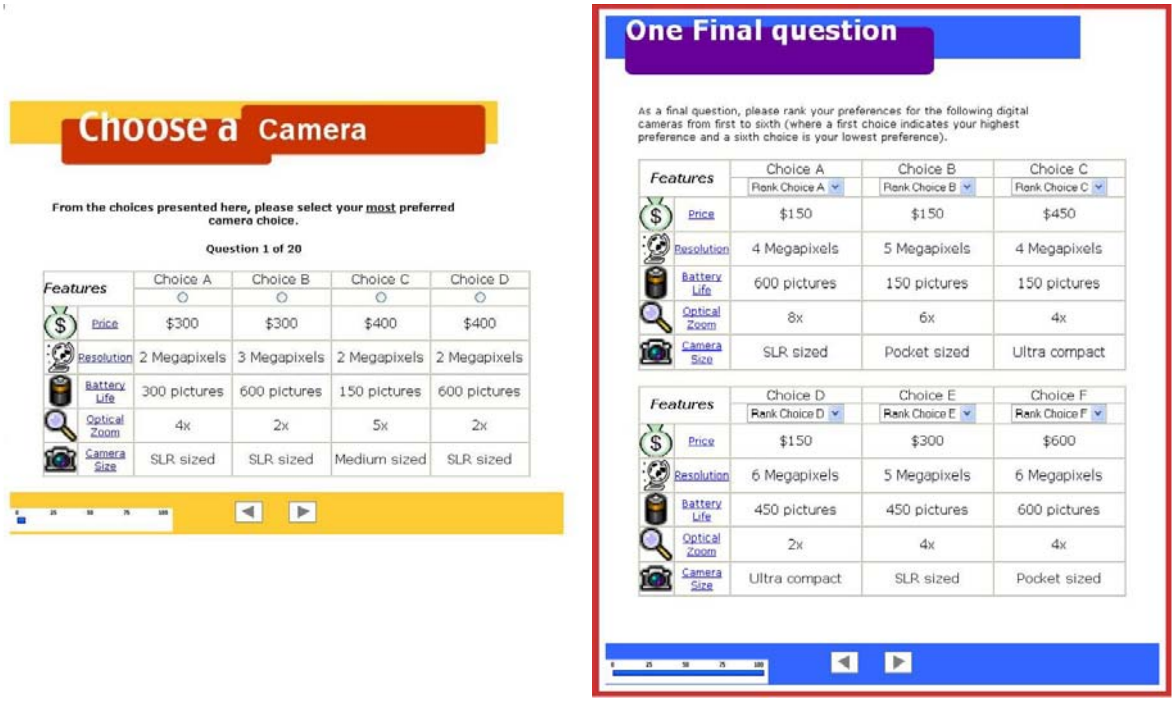
We close by recognizing other research in the area of robust questionnaire design, which was conducted in parallel with the present research. Toubia, Hauser and Garcia (2005) proposed a non-deterministic generalization of the polyhedral method of Toubia et al. (2004). Their approach captures response error by using mixtures of polyhedra to represent posterior beliefs on the partworths.

Magnitude	Heterogeneity	Design	AC estimation	RN estimation	HB estimation
High	Low	Orthogonal	0.731	0.779	0.506
		Customized	0.641	0.601	0.394
		Polyhedral	**0.474**	0.435	0.388
		RN	**0.475**	**0.428**	**0.376**
Low	Low	Orthogonal	**0.801**	0.889	**0.765**
		Customized	0.886	0.914	0.877
		Polyhedral	0.972	0.847	0.894
		RN	0.879	**0.811**	0.780
High	High	Orthogonal	0.784	0.765	0.559
		Customized	0.647	0.529	0.362
		Polyhedral	**0.406**	0.411	0.334
		RN	0.413	**0.383**	**0.325**
Low	High	Orthogonal	**0.757**	0.835	0.692
		Customized	0.933	0.844	0.796
		Polyhedral	0.831	0.739	0.746
		RN	0.776	**0.716**	**0.664**

Table 1: Simulation results (RMSE). Bold number indicate best or not significantly different from best at $p < 0.05$ for each (magnitude × heterogeneity × estimation method) combination.

Performance Metric	Comparison
Holdout Hit Rate	POLY ≈ RN > Agg. Cust. ≈ Orthogonal
Choice hit rate	Agg. Cust. ≈ RN > Orthogonal ≈ POLY
Choice correlation	Orthogonal ≈ RN ≈ Agg. Cust. > POLY

Table 2: Comparison of the question selection methods. "≈" indicates that the regression coefficients are not significantly different at the p <0.05 level, ">" indicates that they are.



Choose a Camera

From the choices presented here, please select your most preferred camera choice.

Question 1 of 20

Features		Choice A ○	Choice B ○	Choice C ○	Choice D ○
$	Price	$300	$300	$400	$400
	Resolution	2 Megapixels	3 Megapixels	2 Megapixels	2 Megapixels
	Battery Life	300 pictures	600 pictures	150 pictures	600 pictures
	Optical Zoom	4x	2x	5x	2x
	Camera Size	SLR sized	SLR sized	Medium sized	SLR sized

One Final question

As a final question, please rank your preferences for the following digital cameras from first to sixth (where a first choice indicates your highest preference and a sixth choice is your lowest preference).

Features		Choice A Rank Choice A	Choice B Rank Choice B	Choice C Rank Choice C
$	Price	$150	$150	$450
	Resolution	4 Megapixels	5 Megapixels	4 Megapixels
	Battery Life	600 pictures	150 pictures	150 pictures
	Optical Zoom	8x	6x	4x
	Camera Size	SLR sized	Pocket sized	Ultra compact

Features		Choice D Rank Choice D	Choice E Rank Choice E	Choice F Rank Choice F
$	Price	$150	$300	$600
	Resolution	6 Megapixels	5 Megapixels	6 Megapixels
	Battery Life	450 pictures	450 pictures	600 pictures
	Optical Zoom	2x	4x	4x
	Camera Size	Ultra compact	SLR sized	Pocket sized

Figure 1: Example Screenshots from the Online Experiment. *Left:* choice-based questions. *Right:* External validity ranking task.

References

[1] Allenby, Greg M., Neeraj Arora, and James L. Ginter (1998) ”On the Heterogeneity of Demand,” Journal of Marketing Research, 35, (August) 384–89.

[2] - - - - and Peter E. Rossi (1999) ”Marketing Models of Consumer Heterogeneity”, Journal of Econometrics, 89, March/April, p. 57 – 78.

[3] Andrews, Rick, Asim Ansari, and Imran Currim (2002) ”Hierarchical Bayes versus finite mixture conjoint analysis models: a comparison of fit, prediction, and partworth recovery”, Journal of Marketing Research, 39, p. 87-98, February 2002.

[4] Arora, Neeraj and Joel Huber (2001) ”Improving parameter estimates and model prediction by aggregate customization in choice experiments”, Journal of Consumer Research, Vol. 28, September 2001.

[5] - - - -, Greg Allenby, and James Ginter (1998) “A Hierarchical Bayes Model of Primary and Secondary Demand”, Marketing Science, 17,1, p. 29–44.

[6] Ben-Akiva, Moshe and Steven R. Lerman (1985), ”Discrete Choice Analysis: Theory and Application to Travel Demand”, MIT Press, Cambridge, MA.

[7] Bettman, James R., Mary Frances Luce, and John W. Payne (1998), ”Constructive Consumer Choice Processes”, *Journal of Consumer Research*, Vol. 25 (December).

[8] Carmone, Frank and Arun Jain (1978), “Robustness of Conjoint Analysis: Some Monte Carlo Results” *Journal of Marketing Research*, 15, p. 300-303.

[9] Cortes, Corinna and Vladimir Vapnik (1995), ‘Support Vector Networks’. *Machine Learning* 20, 1–25.

[10] Cui, Dapeng and David Curry (forthcoming), “Predicting Consumer Choice Using Support Vector Machines with Benchmark Comparisons to Multinomial Logit”, *Marketing Science*, (forthcoming – Manuscript 3003).

[11] DeSarbo, Wayne and Asim Ansari (1997) “Representing Heterogeneity in Consumer Response Models” Marketing Letters, 8:3, p. 335-348.

[12] Evgeniou, Theodoros, Massimiliano Pontil, and Tomaso Poggio (2000), “Regularization Networks and Support Vector Machines” *Advances in Computational Mathematics* 13 (2000), p. 1–50.

[13] - - - -, Constantinos Boussios, and Giorgos Zacharia (2005) ”Generalized Robust Conjoint Estimation”, *Marketing Science*, Vol. 24, No. 3.

[14] Gilbride, Timothy J., and Greg M. Allenby (2004), ”A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules”, *Marketing Science*, Vol. 23, No. 3.

[15] Hauser, John R., Gerald Tellis, and Abbie Griffin (2005), ”Research on Innovation: A Review and Agenda for Marketing Science”, forthcoming, *Marketing Science*.

[16] - - - - and Olivier Toubia (2005), "The Impact of Endogeneity and Utility Balance in Conjoint Analysis", *Marketing Science*, Vol. 24, No. 3.

[17] Huber, Joel, and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs" *Journal of Marketing Research*, 33, p. 307-317.

[18] Johnson, Richard (1987), "Accuracy of utility estimation in ACA, working paper, Sawtooth software, Sequim, WA.

[19] Kanninen, Barbara (2002), "Optimal Design for Multinomial Choice Experiments," *Journal of Marketing Research*, 36 (May), 214–227.

[20] Kohli, Rajeev, and Kamel Jedidi (2005), "Representation and Inference of Lexicographic Preference models and their Variants", working paper, Columbia Business School, New York, NY.

[21] 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, 173–91.

[22] Newey, Whitney K. and Daniel McFadden (1994), "Large Sample Estimation and Hypothesis Testing", in *Handbook of Econometrics, Volume IV*, Edited by R.F. Engle and D.L. McFadden, Elsevier Science.

[23] Payne, John W., James R. Bettman, and Eric J. Johnson (1988), ”Adaptive Strategy Selection in Decision Making”, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 14, No. 3, 534-552.

[24] - - - -, James R. Bettman, and Eric J. Johnson (1993), ”The Adaptive Decision Maker”, Cambridge: Cambridge University Press.

[25] Rifkin, Ryan, Gene Yeo, and Tomaso Poggio (2003) “Regularized Least Squares Classification,” *Advances in Learning Theory: Methods, Model and Applications*, NATO Science Series III: Computer and Systems Sciences, Vol. 190, IOS Press, Amsterdam 2003. Edited by Suykens, Horvath, Basu, Micchelli, and Vandewalle.

[26] Sandor, Zsolt, and Michel Wedel (2001), “Designing Conjoint Choice Experiments Using Managers’ Prior Beliefs” *Journal of Marketing Research*, 38, 4, p. 430-444.

[27] Sawtooth Software, Inc. (1996), ”ACA system: Adaptive conjoint analysis, ACA Manual. Sequim, WA.

[28] - - - - (2002), ”ACA 5.0 Technical paper”, Sequim, WA.

[29] Scholkopf, Bernhard, Chris Burges, and Vladimir Vapnik (1996) “Incorporating Invariances in Support Vector Learning Machines”, Artificial Neural Networks — ICANN’96, Springer Lecture Notes in Computer Science, Vol. 1112, Berlin.

[30] Srinivasan, V. (1998), “A Strict Paired Comparison Linear Programming Approach to Nonmetric Conjoint Analysis” Operations Research: Methods, Models and Appli-

cations, Jay E. Aronson and Stanley Zionts (eds), Westport, CT: Quorum Books, p. 97-111.

[31] - - - - and Allan D. Shocker (1973), "Linear Programming Techniques for Multidimensional Analysis of Preferences" *Psychometrica*, 38,3, p. 337-369.

[32] Tikhonov, A. N., and V. Y. Arsenin (1977), *Solutions of Ill-posed Problems.* W. H. Winston, Washington, D.C.

[33] Toubia, Olivier, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation", *Marketing Science*, 22 (3).

[34] - - - -, John R. Hauser, and Duncan I. Simester (2004), "Polyhedral methods for adaptive choice-based conjoint analysis", *Journal of Marketing Research* (forthcoming).

[35] - - - -, John R. Hauser, and Rosanna Garcia (2005), "Generalized Polyhedral methods for adaptive choice-based conjoint analysis: Theory and Application", working paper, Columbia Business School.

[36] Vapnik, Vladimir (1998), *Statistical Learning Theory.* New York: Wiley.

[37] Yee, Michael, Ely Dahan, Jim Orlin and John R. Hauser (2005), "Greedoid-based Non-Compensatory Consideration-then-Choice Inference", working paper, MIT Sloan School of Management, Cambridge, MA.

Appendices

A RN-Based Questionnaire Design

Regularization networks (RN) are defined as:

$$R_n\left(\mathbf{w}\right) = \sum_{i=1}^{n} \left(1 - \left(\mathbf{x}_{i1} - \mathbf{x}_{i2}\right) \cdot \mathbf{w}\right)^2 + \lambda \mathbf{w}^\top \mathbf{w}.$$

Simple linear algebra shows that R_n and its derivatives can be written in matrix form as follows:

$$\begin{aligned} R_n\left(\mathbf{w}\right) &= \mathbf{w}^\top \left(Z_n^\top Z_n + \lambda \mathbf{I}_p\right) \mathbf{w} - \mathbf{w}^\top Z_n^\top \mathbf{1}_n - \mathbf{1}_n^\top Z_n \mathbf{w} + \mathbf{1}_n^\top \mathbf{1}_n, \\ \nabla R_n\left(\mathbf{w}\right) &= 2\left(Z_n^\top Z_n + \lambda \mathbf{I}_p\right) \mathbf{w} - 2 Z_n^\top \mathbf{1}_n, \\ \nabla^2 R_n\left(\mathbf{w}\right) &= 2\left(Z_n^\top Z_n + \lambda \mathbf{I}_p\right), \end{aligned} \tag{8}$$

where Z_n is the design matrix with rows $Z_{ni} = \left(\mathbf{x}_{i1} - \mathbf{x}_{i2}\right)$ and $\mathbf{1}_n$ is a vector of n 1's.

This implies that the partworth estimate after n questions is given by:

$$\mathbf{w}_n = \left(Z_n^\top Z_n + \lambda \mathbf{I}_p\right)^{-1} Z_n^\top \mathbf{1}_n,$$

and that the direction selected to form the next question corresponds to the eigenvector associated with the smallest positive eigenvalue of the matrix:

$$\left(\mathbf{I}_p - \frac{\mathbf{w}_n \mathbf{w}_n^\top}{\mathbf{w}_n^\top \mathbf{w}_n}\right) \left(Z_n^\top Z_n + \lambda \mathbf{I}_p\right)$$

B Designing Profiles with Discrete Features

We follow the approach of Toubia et al. (2004) which we briefly review here. In particular, to create 4 binary products for a question as done in the experiments, we start from the 2 difference vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$ – corresponding to the smallest and second smallest eigenvalue of the Hessian matrix (4) scaled so that they have square norms of 1 – and our estimate $\mathbf{w}$, and we find a quadrilateral with center w and four corners $\mathbf{c}_1 = \mathbf{w} + \alpha_1 \mathbf{v}_1$, $\mathbf{c}_2 = \mathbf{w} - \beta_1 \mathbf{v}_1$, $\mathbf{c}_3 = \mathbf{w} + \alpha_2 \mathbf{v}_2$, $\mathbf{c}_4 = \mathbf{w} - \beta_2 \mathbf{v}_2$. The α's and $\beta's$ are chosen as the maximum positive real numbers for which the corners are *consistent* with the data points, e.g., $\alpha_1 = \max\{\alpha : \mathbf{z}_i \cdot (\mathbf{w} + \alpha_1 \mathbf{v}_1) \geq 0, \forall i\}$, where the $\mathbf{z}_i$'s are the profile differences from the previous questions. We exclude data points which are misclassified by our estimate, that is, for which $\mathbf{z}_i \cdot \mathbf{w} < 0$.

Having obtained $\{\mathbf{c}_1, \ldots, \mathbf{c}_4\}$, we then find a binary vector $\mathbf{b}_i$ for each $\mathbf{c}_i$. To this purpose we use a knapsack problem. We pick a random budget constraint M, and solve the following problem for each i: maximize $\mathbf{b}_i \cdot \mathbf{c}_i$ subject to the constraint that $\mathbf{b}_i \cdot \mathbf{w} \leq M$. If all the resulting $\mathbf{b}_i$'s are distinct, we use these vectors as our four profiles for the next question. If they are not, we draw another M and repeat the procedure up to k times (in our simulations and experiment k was set to 10, as in Toubia et al., 2004). If the profiles are still not distinct after k draws of M, we simply use the nondistinct set of $\mathbf{b}_i$'s as our question set.

C Polyhedral Methods as a Limit Case

C.1 Polyhedral Estimation Method

Interestingly, Analytic Center (AC) estimation (Toubia et al., 2003; 2004) can be seen as a limit case of the statistical learning theory approach.

AC estimation can be a viewed as a 2-stage procedure. The first step consists in maximizing fit by solving the following problem:

$$\delta^\star = \max_{\mathbf{w}} \min_{i} \{\mathbf{z}_i \cdot \mathbf{w}\},$$

In the case in which the feasible space is non-empty, $\delta^\star$ is simply 0. The second stage can be viewed as minimizing complexity. Indeed, as noted by Evgeniou et al. (2005), finding the analytic center of a polyhedron is a version of complexity control. The exact formulation is as follows:

$$\mathbf{min}_{\ \mathbf{w}} \quad -\textstyle\sum_{i=1}^{n} \ln(\mathbf{z}_i \cdot \mathbf{w} + \delta^\star) - \sum_{i=1}^{p} \ln(w_i) \tag{9}$$

Subject to

$$\mathbf{1}_p \cdot \mathbf{w} = 100,$$

This 2-stage procedure may be viewed as a limit case of the following 1-stage estimation procedure:

$$\mathbf{min}_{\mathbf{w},\delta} \quad \frac{1}{\lambda}\delta\theta\left(\delta\right) - \sum_{i=1}^{n} \ln\left(\mathbf{z}_i \cdot \mathbf{w} + \delta\right) - \sum_{i=1}^{p} \ln\left(w_i\right) \tag{10}$$

Subject to

$$\mathbf{1}_p \cdot \mathbf{w} = 100,$$

A positive value of λ would ensure a trade off between fit and complexity with fit measured by $\delta\theta\left(\delta\right)$ and complexity by $-\sum_{i=1}^{n} \ln\left(\mathbf{z}_i \cdot \mathbf{w} + \delta\right) - \sum_{i=1}^{p} \ln\left(w_i\right)$. (Note that this problem may not be solvable efficiently.) However when λ goes to 0, the relative weight on fit goes to $+\infty$. The δ solution of (10) converges to $\delta^\star$, and the $\mathbf{w}$ solution of (10) converges to the solution of (9).

C.2 Polyhedral Questionnaire Design

When designing questions, Toubia et al (2004)'s polyhedral method is such that the space of partworths solutions is always feasible – the estimated partworths fit all previous responses and estimation is equivalent to (9) with $\delta^\star = 0$. If we were to use the Hessian-based approach developed in this paper in conjunction with AC estimation, the Hessian of (9) would be:

$$\nabla^2 R_n = Z_n^\top (D_n)^{-2} Z_n + W_n^{-2} \tag{11}$$

where D_n is the $n \times n$ diagonal matrix with entry $D_n(i,i) = (\mathbf{x}_{i1} - \mathbf{x}_{i2}) \cdot \mathbf{w}_n$, and W_n is the $p \times p$ diagonal matrix with $W_n(i) = w_{ni}$ (the i^{th} element of $\mathbf{w}_n$). The next question $\mathbf{z}_{n+1}$ would then be defined by the eigenvector associated with the smallest positive eigenvalue of:

$$\left(\mathbf{I}_p - \frac{\mathbf{w}_n \mathbf{w}_n^\top}{\mathbf{w}_n^\top \mathbf{w}_n}\right) (Z_n^\top (D_n)^{-2} Z_n + Z_n^{-2}) \tag{12}$$

Instead, Toubia et al (2004) use the eigenvector associated with the smallest positive eigenvalue of:

$$U_n^{-2} - V_n^\top (V_n V_n^\top)^{-1} V_n U_n^{-2} \tag{13}$$

where U is the $(p+n)$ diagonal matrix $[W_n \;\; 0_{p \times n};\; 0_{n \times p} \;\; D_n]$ and V_n is the $(n+1) \times (p+n)$ matrix $[X_n \;\; -\mathbf{I}_n;\; \mathbf{1}_p \; 0_n]$ where $\mathbf{1}_p$ is a row vector of p ones, 0_n is a row vector of n zeros, and $0_{p \times n}$ a $(p \times n)$ matrix of zeros. The matrices U and V incorporate slack variables that enforce the scaling constraint $\mathbf{1}_p \cdot \mathbf{w} = 100$ as well as the positivity constraints imposed by the choices. Geometrically, this eigenvector corresponds to the longest axis of an ellipsoid that approximates the polyhedron defined by: $\{\mathbf{w} \;\mid\; V_n \cdot \mathbf{w} = [0_n; 100],\; \mathbf{w} \geq 0\}$. Utility balance is achieved through the Knapsack problem that translates the real-valued longest axes into binary profile vectors.

We see that although the two question selection methods are not exactly equivalent, they are similar in spirit. In particular, they both rely on the identification of a "high uncertainty" direction: longest axis versus flattest direction of the loss function. In both cases the optimal direction is defined by the smallest positive eigenvector of a matrix describing the space of partworth solutions.

D Results of the online experiment based on 16 questions

	Hold Out Hit Rate			*Choice Hit Rate*			*Choice Correlation*		
Design	AC	RN	HB	AC	RN	HB	AC	RN	HB
Orthogonal	2.13	2.15	2.56	0.440	0.504	0.592	0.506	0.560	0.644
Customized	2.14	1.98	2.50	0.456	0.512	0.560	0.504	0.502	0.594
Polyhedral	2.21	2.17	2.50	0.440	0.440	0.520	0.484	0.495	0.551
RN	2.34	2.36	2.51	0.512	0.464	0.552	0.528	0.519	0.601

Table 3: Results of the online experiment based on 16 questions. 6 pairwise comparisons of question selection methods are possible within each 9 performance metric × estimation method column. Out of all possible comparisons, 2 are significant: RN performs significantly better than aggregate customization in the holdout hit rates × RN combination (second column), and Orthogonal performs significantly better than POLY in the choice correlation × HB combination (last column).