Optimal Product Design by Sequential Experiments in High Dimensions

by

Mingyu Joo
Ohio State University
joo.85@osu.edu

Michael L. Thompson
The Procter and Gamble Company
thompson.ml@pg.com

Greg M. Allenby
Ohio State University
allenby.1@osu.edu

March 26, 2018

The authors thank Eric Anderson, the associate editor, and two anonymous reviewers for their valuable and insightful comments. They also thank Angela Dean, Diane Farris, Sang Pil Han, Scott Hayes, David Hengehold, Uma Karmarkar, Dong Soo Kim, Lan Luo, Bryan Orme, Adam Smith, Olivier Toubia, Ken Wilbur and seminar participants at Arizona State University, Brigham Young University, Erasmus University, UC San Diego, INFORMS Marketing Science, Sawtooth Software conference, Theory and Practice in Marketing, and UTD Frank M. Bass conference for their helpful comments and discussions.

Optimal Product Design by Sequential Experiments in High Dimensions

Abstract

The identification of optimal product and package designs is challenged when attributes and their levels interact. Firms recognize this by testing trial products and designs prior to launch where the effects of interactions are revealed. A difficulty in conducting analysis for product design is dealing with the high dimensionality of the design space and the selection of promising product configurations for testing. We propose an experimental criterion for efficiently testing product profiles with high demand potential in sequential experiments. The criterion is based on the expected improvement in market share of a design beyond the current best alternative. We also incorporate a stochastic search variable selection method to selectively estimate relevant interactions among the attributes. A validation experiment confirms that our proposed method leads to improved design concepts in a high-dimensional space compared to alternative methods.

Keywords: Design Criterion, Expected Improvement, Interaction Effects, Stochastic Search Variable Selection

1 Introduction

An optimal product or package design is one having the most preferred combination of attributes from a feasible set of candidates (e.g., color and font combinations on product packages, brand logos, etc.) The presence of interactive effects among attributes and their levels challenges the identification of optimal configurations because they typically lead to a dramatic increase in the dimensionality of the design space. Consider, for example, the optimal design for a container of liquid detergent, where the color of the bottle, color of the cap and color of lettering are being evaluated. The dimensionality of the design space for just 10 colors for each is at least 300 when considering the two-way interactions alone. As such, product and package design experiments quickly become unwieldy for even the simplest settings when interactive effects are thought to be present.[1]

Design experiments for product optimization consist of two components: the selection of design points to learn about consumer preferences, and models for relating preferences to product features and attributes. Commonly used design criteria for evaluating design points, such as D- and A-efficiency measures, have focused on learning about preferences for all product features, including those associated with configurations that consumers don't want. These design points, however, are not aligned with the goal of identifying product and package configurations that maximize demand. The allocation of design points to the overall range of attribute-levels sacrifices potential opportunities to learn more about parameters associated with regions of high preference.

We propose a new design criterion for choice-based conjoint analysis that adaptively selects design points to improve the aggregate market shares of product and package designs. A linearized aggregate share model is used to identify important product attributes while minimizing the effects of heterogeneous consumer preferences. The share model directly operationalizes a firm's goal function in product design problems and minimizes the impact of respondent-level errors in detecting interactions. In addition, we also show that Bayesian variable selection methods (SSVS, George and McCulloch 1993) are effective for managing the high dimensionality of

[1] A well known example in practice is Google's testing of 41 shades of blue in their logo design.

our proposed model that allows for interactive effects.

The proposed design criterion systematically balances between exploitation of high performance products and exploration of high potential products by selecting design points, or product profiles, with high uncertainty in high preference regions of the design space. The selected product profiles are most likely to outperform the current best design in the subsequent rounds of a sequential experiment. In addition, the design criterion is evaluated with full-product profiles so that it can capture interactive effects among the attribute-levels. Data from a national sample of consumers is used to identify an optimal package design for a nationally distributed product, and we demonstrate that it outperforms other package designs identified by competing methods. The empirical application confirms that the proposed criterion tends to include more combinations with important interactive effects than expected.

The organization of the paper is as follows. Section 2 discusses the relationship between the proposed criteria of expected improvement and the existing methods of adaptive experiments. Section 3 develops the optimal sequential search criteria, illustrates the SSVS estimation, and presents a simulation study to compare the performance of the proposed framework to that of an alternative method. Section 4 presents an empirical application of the proposed framework to a design project in practice. Section 5 discusses the results using the national sample, and presents a validation experiment. Concluding comments are provided in Section 6.

2 Relationship to prior literature

We balance the goals of exploration and exploitation in a product design experiment by selecting design points that are expected to improve the product's aggregate market share. Our goal is similar that encountered in the context of optimal search, where an analyst sequentially acquires information believed to be maximally informative about an outcome of interest. We measure the value of a design point relative to the best performing design found in earlier rounds of an experiment. Prioritizing profiles with the highest expected improvement in aggregate share takes an outcome of importance to the firm (sales) as the design criteria rather than using a statistical criterion, such as the determinant of the covariance matrix of model coefficients. In

doing so, the analyst focuses on highly preferred combinations of product features that leads to more extensive comparisons of promising interactive effects.

Our expected improvement criterion is closely related to Weitzman (1979)'s optimal search rule by relating the posterior predictive performance of a design to the observed best outcome. Weitzman (1979) showed that a firm can most efficiently search for the best alternative in an R&D project by pursuing options one-by-one in descending order of reservation prices, favoring those with higher reward from a known, invariant performance distribution. We show that our evaluation of the expected share improvement presents an equivalent rank order to the Weitzman's reservation price rule in the absence of costs. In addition, we do not rely on analytical solutions nor employ option values (i.e., reservation prices) as a means of determining an optimal stopping rule.

Our model contributes to the literature in question-selection methods in conjoint experiments. Toubia et al. (2003, 2004) develop a method for adaptive conjoint analysis that selects the next question based on the combination of coefficients currently measured with least certainty, where the linear inequalities are geometrically presented as an uncertainty polyhedron. Toubia et al. (2007) generalize the method to the domains with high response errors by adding a probabilistic structure to the region of combination of coefficients. Sauré and Vielma (2017) propose an improvement on the polyhedral method by using an uncertainty ellipsoid that represents a credibility region for the posterior distributions of coefficients. Huang and Luo (2016) suggest an algorithm of selecting adaptive questionnaire that maximally reduces the feasible region of coefficients using support vector machine learning. Finally, an algorithm by Dzyabura and Hauser (2011) selects questions that minimize uncertainty in respondents' decision heuristics in order to distinguish between non-compensatory screening and preference ordering. These approaches all aim to reduce overall parameter uncertainty can be viewed as adaptive implementation of D-efficiency (Toubia et al. 2004, Sauré and Vielma 2017). In contrast, our criterion aims to learn more about highly preferred attribute combinations with less focus on profiles in the low utility region.

The statistics literature has examined the efficiency of the *expected improvement* criterion

relative to *expectation maximization* in the optimization of a goal function (Jones et al. 1998).[2] The expected improvement criterion can be viewed as a one-step-ahead optimality criterion by choosing profiles that are expected to outperform the current optimal profile. Although it does not provide an exact optimization of a goal function, it is sufficient to determine the point of the next observation (Mockus 1994, Schonlau et al. 1997). The criterion is also statistically shown to require relatively small number of design points for global optimization by balancing between global search (exploration) and local search (exploitation) (Schonlau et al. 1997, 1998). The literature supports the notion that the expected improvement criterion offers data points to enhance efficiency of the second step of decision making – global optimization of a goal function.

The proposed framework is also related to machine-learning and artificial-intelligence algorithms in decision support systems. A stream of research suggests to select a query that maximizes expected value of information in statistical measures such as the likelihood of a model (Cavagnaro et al. 2010) or the expected performance of all alternatives in a query set (Viappiani and Boutilier 2010). Another stream proposes to select a query that minimizes regret or cost, when making decisions under strict uncertainty (e.g., Wang and Boutilier 2003). An independent yet related methodology is the multi-armed bandit (MAB), which simultaneously considers maximum performance and minimum regret (Schwartz et al. 2016). It is widely used in online field experiments for digital content optimization (e.g., Brezzi and Lai 2002, Scott 2010, Schwartz et al. 2016) to balance between profits during the experimental phase and the optimal outcome. While the criteria in these literatures have not been developed for, and applied to, the question-selection problems, their goal-directed nature is conceptually similar.

In sum, we propose a goal-directed question-selection method in choice-based conjoint analysis for design optimization problem.[3] While the extant methods focus on uncertainty reduction for precise preference estimation, the proposed selection criterion jointly considers preference

[2]The literature mainly focuses on the engineering context including automotive and semiconductor industries, so the explanatory variables are strictly ordinal and continuous with objective performance variation (e.g., engine size).

[3]Adaptive choice-based conjoint analysis (ACBC) proposed by Sawtooth software also aims to sequentially present highly preferred profiles. It first determines an approximate region that a respondent prefers by build-your-own profile task (BYO), then asks unacceptables and must-haves. The knowledge from these tasks identifies a consideration set of a respondent, and the criterion offers D-efficient design within this consideration set. Therefore, it is more of a consideration-set driven criterion, rather than a goal-driven one. For more details, see https://www.sawtoothsoftware.com.

elicitation (i.e., exploration) and goal achievement (i.e., exploitation) in high dimensions.

3 Model Development

We develop our model and design criterion in the context of an aggregate choice experiment in which respondents are asked to identify their favorite design concept from among n product profiles and an outside option. The outside option can be either a benchmark design that is common across questions, or a no-choice option given the n product profiles. Each respondent receives Q questions, so $(nQ + 1)$ product profiles or design concepts including the outside option are evaluated per respondent.

We assume that the data generating process follows a multinomial logit choice model for the $(n + 1)$ options in each question. Aggregation across respondents yields choice shares of $(n + 1)$ alternatives in each question $q \in Q$. The choice share of a design concept j is given by S_{jq} and that of the outside option is given by S_{0q}. The choice shares represent standard logit choice probabilities of respondents at the aggregate level as follows:

$$S_{jq} = \frac{\exp\left(u\left(X_j;\boldsymbol{\beta}\right)\right)}{1+\sum_{j'\in J_q}\exp\left(u\left(X_{j'};\boldsymbol{\beta}\right)\right)} \text{ and } S_{0q} = \frac{1}{1+\sum_{j'\in J_q}\exp\left(u\left(X_{j'};\boldsymbol{\beta}\right)\right)},$$

where the utility of a design concept j, $u\left(X_j;\boldsymbol{\beta}\right)$, is determined by a linear combination of the design attributes X_j and a vector of aggregate partworth preference parameters $\boldsymbol{\beta}$. J_q is a set of n alternatives evaluated in question q, and the utility of the outside option is normalized to be zero. Taking the log-odds ratio of S_{jq} and S_{0q} in question q linearizes the logit choice model (Allenby and Rossi 1991, Berry 1994), with the aggregate log shares given by:

$$Y_{jq} = \ln S_{jq} - \ln S_{0q} = u\left(X_j;\boldsymbol{\beta}\right) = X_j\boldsymbol{\beta} + \xi_{jq},\ \ \xi_{jq}\sim N\left(0,\sigma^2\right) \tag{1}$$

where ξ_{jq} is a normal error term specific to a product profile j in question q.

Use of the share model aligns with the objective of product development - to design products that maximize market share. A benefit of using a share model is that the aggregate data are measured with greater certainty and have greater likelihood of detecting potential interactions. One might be concerned that the model does not incorporate respondents' heterogeneity. How-

ever, if the goal is to predict aggregate responses, the coefficients of the aggregate model can be shown to recover well the average of heterogeneous individual-level coefficients because of the log-linear model structure.[4] Thus, the estimates of Equation (1) are sufficient for predicting market-level responses despite the absence of heterogeneity.

Our goal in the design optimization problem is to identify product profiles j with the greatest chance of market success with high $u(X_j; \boldsymbol{\beta})$. To achieve this goal, we propose a question-selection criterion that conditions on the best outcome to date to identify product profiles with the highest expected improvement in the market share. In contrast to the extant uncertainty reduction methods (e.g., Toubia et al. 2004), uncertainty is unevenly reduced among parameters as the sample size increases and learning is concentrated on the highly valued product profiles. In addition, our measure focuses on exceeding the value of the best outcome to date without considering the product profiles from whence that outcome originates, which balances between exploitation and exploration. We couple our goal-directed criterion with an SSVS prior to efficiently navigate the high dimensional design space.

3.1 Goal-directed criterion for sequential experiments

Suppose that the number of rounds in the sequential design experiment is T and the number of questions per round is Q. As each respondent chooses the most preferred design concept per question out of n product profiles and one common benchmark design (outside option), each round $t\,(\in T)$ evaluates nQ product profiles and one common outside option. The outside option is predetermined by the researcher using domain knowledge. It can be the current product design on the market or a basic 'vanilla' design concept for the same product. Inclusion of the common outside option across all questions makes preference measures robust to different combinations of n choice options.

The first nQ product profiles in round 1 can be selected either randomly or by any subset of classical experimental criteria.[5] The log-odds ratio between the aggregated market shares of nQ product profiles and the outside option in each question constructs nQ data points for

[4] See Appendix D for empirical simulation tests for this claim.

[5] We do not apply the adaptive criteria in selection of the initial set of candidates, because the partworth parameters are simply priors before the first round.

the share model in Equation (1). The partworth preference parameters $\boldsymbol{\beta}$ are estimated at the end of every round using the cumulative data collected until the current round. The next nQ product profiles in the subsequent round 2 are determined by the selection criteria to maximize the expected improvement relative to the current best outcome of the experiment.

The proposed selection rule evaluates the expected improvement in market share by searching for a new product profile in the next round. The expected improvement is defined as the upper-tail expectation of a new product profile's preference distribution in the range that it outperforms the current best design. As consumer preferences are unknown before the experiment is run, it evaluates the predictive distributions of unevaluated design concepts based on the parameter estimates using the partial cumulative data. A product profile with the highest value of expected improvement is the most likely to outperform the best observed outcome so far. It prioritizes such product profiles with high expected improvement for the next round of the experiment, as it indicates an evaluation of potential or one-step-ahead optimality.

The expected improvement is conceptualized from the expected return to search for a new product profile in the presence of the current best knowledge. The expected return of testing an additional design concept j in round $(t+1)$ with the best known outcome z_t in round t is given by

$$V\left(X_j\right) = \int_{-\infty}^{z_t} z_t dF\left(u; X_j, \boldsymbol{\beta}\right) + \int_{z_t}^{\infty} u\left(X_j; \boldsymbol{\beta}\right) dF\left(u; X_j, \boldsymbol{\beta}\right),\ j \in J \setminus J_t \tag{2}$$

where $F\left(u; X_j, \boldsymbol{\beta}\right)$ is the cumulative density function of the predictive utility distribution of a design concept j,[6] J denotes a set of potential combinations of all attribute levels constructing a complete list of candidate product profiles, and J_t denotes a set of all product profiles evaluated up to round t. The current best outcome z_t is formally defined as

$$z_t = \max\ u_j\left(X_j; \boldsymbol{\beta}\right),\ j \in J_t \tag{3}$$

We note that z_t is an observed value obtained by market shares up to round t as in Equation (1).

[6] As u_j is defined as $X_j\boldsymbol{\beta}$ in Equation (1), the probability density function of u_j is obtained by the posterior distribution of $X_j\boldsymbol{\beta}$.

The first element of the two additive terms in $V(X_j)$ means that the expected outcome is still z_t when the new design concept j turns out to be worse than z_t. The second element stands for the expected outcome when the new design concept j performs better than z_t, which represents the expected improvement by testing j in the next round. This conditional expectation is a better reflection of the goal to find the best product profile than an evaluation of the expected value of $u(X_j;\boldsymbol{\beta})$, because the experiment still keeps the known outcome of z_t and abandon j when $u(X_j;\boldsymbol{\beta})$ is lower than z_t.[7]

It is straightforward to show that the comparison of $V(X_j)$ among different design concepts j is equivalent to the comparison of the upper tails only (i.e., expected improvement). That is, if

$$\int_{z_t}^{\infty} u(X_j;\boldsymbol{\beta})\, dF(u;X_j,\boldsymbol{\beta}) > \int_{z_t}^{\infty} u\left(X_{j'};\boldsymbol{\beta}\right) dF\left(u;X_{j'},\boldsymbol{\beta}\right)$$

then $V(X_j) > V\left(X_{j'}\right)$ for all j and j' where $j' \neq j$. The proof of this equivalence is presented in Appendix A. Therefore, the expected improvement $V'(X_j)$ represents the rank order of product profiles that are not yet tested in terms of expected return to search for additional design points:

$$V'(X_j) = \int_{z_t}^{\infty} u(X_j;\boldsymbol{\beta})\, dF(u;X_j,\boldsymbol{\beta}),\ j \in J \setminus J_t \tag{4}$$

As $V'(X_j)$ is obtained as an expected value of u_j in the range that is higher than z_t, it returns a scalar value, not a distribution, that is easily comparable across different product profiles.

The nQ product profiles with the highest $V'(X_j)$ are selected for testing in round $(t+1)$ among the options that are not evaluated so far (i.e., $j \in J \setminus J_t$). The partworth preference parameters $\boldsymbol{\beta}$ are updated after round $(t+1)$ using all the cumulative data collected from rounds 1 to $(t+1)$. The same procedure is iterated until round T. The number of rounds T can be determined based on the required data points for stable parameter estimates given the dimensionality of attributes and/or a research project's budget consideration. The proposed criterion is considered to be as an aggressive one, preferring high variance in the predictive utility distribution of design concepts. However, it balances with exploitation of high performance

[7] For further discussion, see Weitzman (1979).

region by focusing on upper tails, which does not waste resources for random exploration.

The sequential experiment determines explanatory variables in each round based on the observed dependent variables in the previous rounds, leading to a concern of potential endogeneity biases in the partworth parameter estimates. We note that the endogeneity coming from the adaptive selection criteria is ignorable under our Bayesian inference according to the likelihood principle (Liu et al. 2007). The selection of product profiles and corresponding explanatory variables in round $(t+1)$ is deterministic, given all observed dependent variables in rounds 1 to t. Therefore, conditional on all the observed data, the selection criterion does not affect the likelihood, and parameter estimates are consistent.

The expected improvement criterion $V'(X_j)$ is determined by $u(X_j;\boldsymbol{\beta})$ and its cumulative density function $F(u;X_j,\boldsymbol{\beta})$, not directly by individual partworth parameters $\boldsymbol{\beta}$. So, it is not dependent upon specific modeling choice of $u(X_j;\boldsymbol{\beta})$. Therefore, the criterion can flexibly accommodate various expansion in the model specification, such as inclusion of interactive effects where $X_j = [X_{j1} \quad X_{j2}]$ with main effects X_{j1} and higher order interactions X_{j2}. As $V'(X_j)$ returns a scalar value regardless of the dimensionality of X_j, the proposed criterion can be easily applied to problems with high dimensional X_j.

3.2 Selection of relevant variables

A stochastic search variable selection (SSVS) method is used to selectively estimate relevant covariates in both main and interaction effects (George and McCulloch 1993; George and McCulloch 1997; Gilbride et al. 2006). The dimension of covariate space in the design problem is extremely high especially with interaction effects, and not all interaction terms affect respondents' preferences leading to a very sparse parameter space. The inclusion of irrelevant covariates taxes the ability of the model to conduct accurate inferences by leading to a higher dimensional model than the number of observations. The SSVS method effectively collapses the irrelevant covariates toward zero and identifies covariates that are used in respondents' evaluation of the design concepts.

The relationship between the dependent variable and covariates in the normal linear model

in Equation (1) is given by:

$$f\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2\right)=N\left(\boldsymbol{X}\boldsymbol{\beta},\sigma^2 I\right)$$

where $\boldsymbol{Y}$ is a $nQt\times 1$ vector of all dependent variables at the end of round t, $\boldsymbol{X}$ is $nQt\times p$ matrix of covariates, and the variance σ^2 is a scalar value with a prior distribution of $InvertedGamma\left(\kappa,\kappa\psi\right)$. A latent variable γ_i represents the selection of a particular individual parameter β_i through its prior distribution as follows:

$$\pi\left(\beta_i|\gamma_i\right)=\left(1-\gamma_i\right)N\left(0,\tau_i^2\right)+\gamma_i N\left(0,c_i^2\tau_i^2\right)$$
$$P\left(\gamma_i=1\right)=1-P\left(\gamma_i=0\right)=p_i$$

where $\gamma_i=1$ if β_i is selected with probability of p_i, and $\gamma_i=0$ if β_i is not selected with probability of $(1-p_i)$. τ_i are small positive values that shrink β_i toward zero when the variable is not selected, and c_i are large positive values to estimate non-zero β_i.[8] The effect of the mixture prior is that β_i is drawn from a mass concentrated around zero when it is not selected.

Incorporating the mixture distribution, the multivariate normal prior for $\boldsymbol{\beta}$ is given by:

$$\pi\left(\boldsymbol{\beta}|\gamma\right)=N\left(0,\boldsymbol{D}_\gamma\boldsymbol{R}\boldsymbol{D}_\gamma\right)$$

$\gamma=\left(\gamma_1,...,\gamma_p\right)'$ and $\boldsymbol{R}$ is prior correlation matrix. $\boldsymbol{D}_\gamma$ summarizes variable selection as follows:

$$\boldsymbol{D}_\gamma\equiv\text{diag}[a_1\tau_1,...,a_p\tau_p]$$

where $a_i=1$ if $\gamma_i=0$ and $a_i=c_i$ if $\gamma_i=1$. The model selection parameters γ and partworth preference parameters $\boldsymbol{\beta}$ are simultaneously estimated by Gibbs sampler as presented in Appendix B.[9]

The SSVS method estimates the partworth preference parameters $\boldsymbol{\beta^{(t)}}$ at the end of round t of data collection. The cumulative data points from round 1 up to round t are used for estimation at the end of the round. The posterior distributions of $\boldsymbol{\beta^{(t)}}$ from SSVS estimation after round t

[8] See George and McCulloch (1993) for selection of c_i.

[9] The SSVS estimation code used in the paper is available from the authors upon request. Free software packages are also available in R.

construct predictive distributions of unknown product profiles ($F\left(u; X_j, \boldsymbol{\beta^{(t)}}\right)$, $J \setminus J_t$) and their upper-tail expectations ($V'(X_j)$), where $V'(X_j)$ determines the set of product profiles to test in round $(t+1)$.

The SSVS method is particularly useful in detecting irrelevant interaction effects to efficiently manage dimensionality with small number of observations, and improves precision of estimates with tighter credible intervals relative to competing estimation methods as shown in Appendix C. We note that the SSVS estimation is purely based on the likelihood and independent of the selection criteria, so it does not restrict the order of interactions in the model specification. If needed, researchers may adopt an alternative estimation method that best suits their model specification along with the proposed question-selection criteria.

3.3 Simulation study

A simulation study is conducted to evaluate how efficiently the proposed modeling framework identifies the true best design profile. We first evaluate the predictive performance of the proposed selection criteria using SSVS estimation, perform sensitivity tests with various numbers of respondents and questions per round, and then present a benchmark study of an alternative preference elicitation method proposed by Toubia et al. (2004).

3.3.1 Experimental setting

We assume that the number of product profiles for testing in each question is $4\,(=n)$ and simulate choice share data from the true partworth preference parameters. Simulated respondents choose their favorite design concept out of five profiles including the outside option. Se also assume that there are $25\,(=Q)$ questions per round, so each round of the experiment evaluates $100\,(=nQ)$ product profiles in addition to the outside option. The number of rounds is assumed to be $5\,(=T)$, for a total of $500\,(=nQT)$ product profile evaluations plus the outside option. The number of respondents in each round is assumed to be $100\,(=H)$.

Each product profile is a combination of four attributes, and these attributes have design candidates of $(5, 8, 11, 12)$ levels each.[10] The true partworth preference parameters are designed

[10]The number of attributes and levels is set to be similar to the collaborating firm's R&D project. The dimensionality in the current simulation study is higher than typical assumptions in the literature (see e.g.,

in a way that the true best design concept is not a combination of the most preferred individual attribute-levels, such that some two-way interactions among attributes are large enough to affect the most preferred profile. Ranges of true partworth preference parameters ($\boldsymbol{\beta}$) are in between 3 and -4.6, and variances for respondent heterogeneity ($\boldsymbol{\Sigma}$) are one for main effects and 0.5 for interaction effects.[11] Each respondent h's true partworth preference parameters ($\boldsymbol{\beta}_h$) are randomly drawn from the true values with heterogeneity distribution (i.e., $\boldsymbol{\beta}_h \sim N(\boldsymbol{\beta}, \Sigma)$).

The market share of a product profile j in question q is simulated as

$$S_{jq} = \frac{1}{H} \sum_{h=1}^{H} \mathbf{1}\{(X_{jq}\boldsymbol{\beta}_h + \epsilon_{hjq}) = \max_{j' \in J_q} \left(X_{jq}\boldsymbol{\beta}_h + \epsilon_{hj'q}\right)\}, \tag{5}$$

where the indicator function $\mathbf{1}\{\cdot\}$ is one if a profile j presents the highest utility among all profiles j' and outside option shown in question q, and zero otherwise. ϵ_{hjq} are simulated logit error terms. Five product profiles in each question (J_q) include an outside option (i.e., $j = 0$) for all questions q. The common outside option across questions is defined as a combination of one of the candidate attribute levels. The true partworth parameter values of the attribute levels in the outside option is normalized to be zero. Each dependent variable ($\ln S_{jq} - \ln S_{0q}$) for a design concept j is calculated in each question q across H individuals.

The dummy coding of all design attributes leads to 32 main effects and 369 two-way interaction effects. The product design problem includes $5,279$ candidate product profiles. Therefore, the simulation tests the performance of the proposed framework when the number of product profiles tested (500) is similar to the number of covariates (401), and the data used is less than 10% of the entire set. We focus on the two-way interaction effects for the case of a relatively small number of attributes (four). However, it is straightforward to extend the framework by simply adding higher-order interaction terms in Equation (1), as we do not impose any specific assumption to the order of interactions.

Toubia et al. 2004).

[11] The variance-covariance matrix of the true partworth preference parameters $\boldsymbol{\Sigma}$ is assumed to be diagonal.

3.3.2 Predictive performance

Partworth preference parameters are estimated at the end of each round t using cumulative data with (nQt) observations. Using the parameter estimates, we present three predictive performance measures – utility regret, in-sample prediction, and out-of-sample prediction – at the end of every round t. The next set of (nQ) observations for round $(t+1)$ is selected by our proposed criteria in Equation (4). Utility regret presents how much performance loss would have happened if the best product profile predicted by the model is launched on the market relative to the true best product profile. More precisely, we present the following % measure of loss of potential mis-prediction:

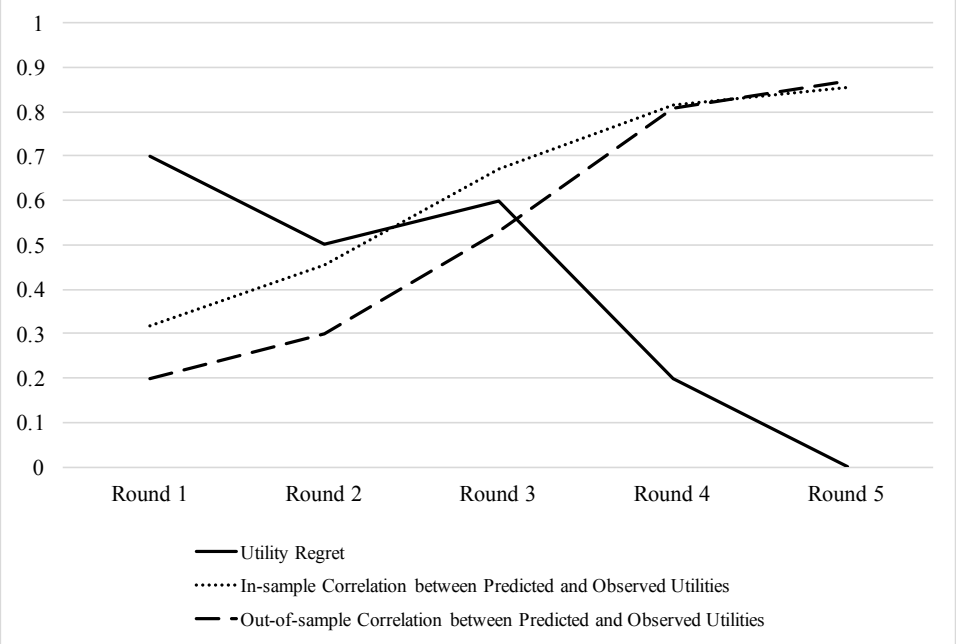
$$\text{Utility Regret} = \frac{\text{(True Utility of } \textit{True} \text{ Best Profile) - (True Utility of } \textit{Predicted} \text{ Best Profile)}}{\text{(True Utility of } \textit{True} \text{ Best Profile)}},$$

where the utility regret measure becomes zero, if a model correctly predicts the true best product profile.

In-sample prediction is measured by the correlation between observed and predicted utilities of (nQt) observations until round t, and out-of-sample prediction is measured by the correlation between observed and predicted utilities of all $5,279$ candidate profiles. The observed utilities in out-of-sample prediction are re-simulated using Equation (1) adding random draws of potential market-level shocks ξ_{jq}. Therefore, observations in the estimation sample are not re-used for out-of-sample validation, and the validation task conceptually replicates firms' R&D practice of testing their knowledge learned from lab studies in actual market.

Figure 1 presents the three performance measures at the end of each round of experiments. Utility regret is high after the initial round testing only 100 profiles, where our question selection criteria were not applied. However, it quickly converges to zero at the end of round 5 after testing additional 400 profiles selected by the proposed criteria. Both in-sample and out-of-sample correlations monotonically increase up to around .85. The results indicate that the proposed selection criteria along with SSVS estimation correctly predicts the best product profile and well predicts overall market share of all $5,279$ candidates after testing a small subset of product profiles.

Figure 1: Predictive performance of the proposed framework



3.3.3 Sensitivity tests

A remaining question may be how efficient the proposed method is, so we performed two additional sensitivity analyses. First, we varied the number of respondents per round (I) to be 50, 25, and 10 to investigate how the stability of market share calculation would affect the predictive performance of the proposed framework with a fixed number of questions per round ($Q = 25$). Next, we varied the number of questions per round (Q) to be 20, 15, 10 to investigate how the number of observations tested in each round affect the predictive performance with a fixed number of respondents per round ($H = 100$).

Figure 2 presents utility regret, in-sample prediction, and out-of-sample prediction after all five rounds of experiments with different numbers of respondents answering 25 questions per round. The proposed framework correctly predicts the best product profile until the number of respondents are as small as 25, but utility regret is increased to 0.2 when only 10 respondents are used to simulate market share observations. Both in-sample and out-of-sample utility predictions decrease with the decreasing number of respondents, but they are still around 0.7 under the worst case. The results indicate that the proposed framework predicts the true best profile with a small number of respondents (25 per round, and 100 in total) given dimensionality of the problem.

Figure 2: Predicted performance by the number of respondents (H, with $Q = 25$)

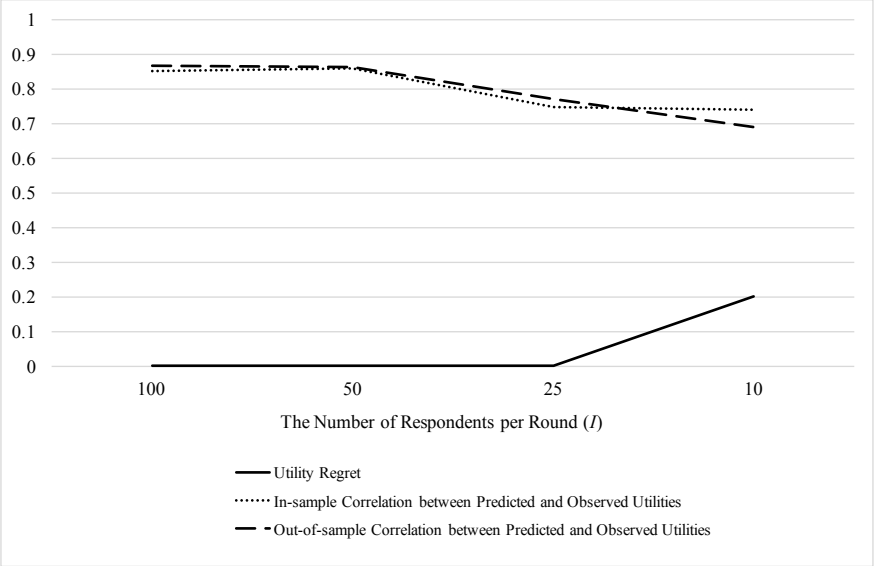


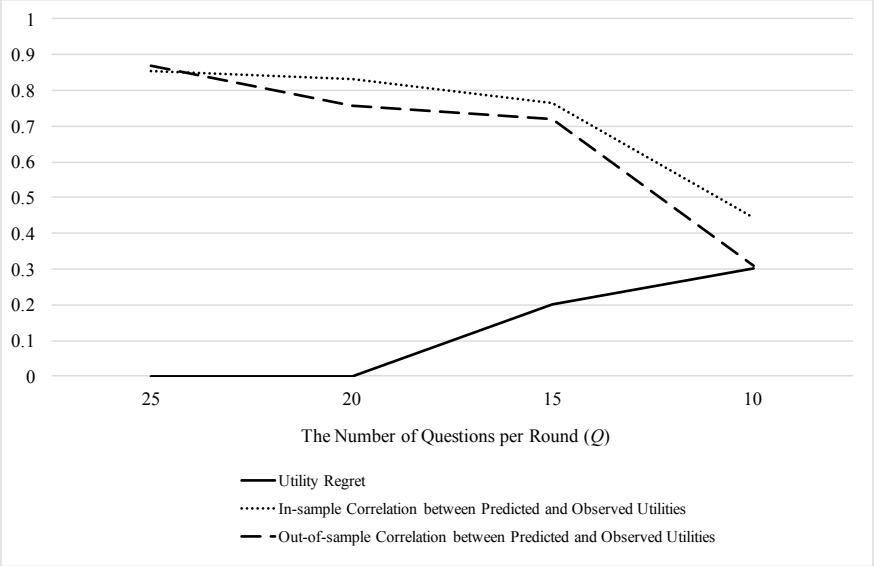
Figure 3 presents the measures of predictive performance after all five rounds with different numbers of questions and 100 respondents per round. The proposed framework correctly predicts the best product profile with 20 questions per round, where the number of profiles tested, 400 =(20 questions x 4 profiles per question x 5 rounds), is smaller than the number of parameters to be estimated (401). Utility regret starts to increase with 15 and 10 questions per round due to small number of observations with 300 and 200 profiles, respectively. In-sample and out-of-sample predictions are reasonably stable (over 0.7) until 15 questions per round, but quickly drop to below 0.5 with 10 questions per round. The results indicate that the predictions are reliable with about 20 questions per round under the current dimension of product profiles.

In sum, the sensitivity analyses show that the proposed framework reliably predicts the true best product profile with a small number of respondents (25 per round) or a small number of questions (20 per round) under the current dimension of the problem, which may guide practitioners' trade-offs between budget consideration and performance of R&D projects.

3.3.4 Benchmark study

Recent advances in preference elicitation methods in choice-based conjoint analysis focus on efficiently learning respondent-level preferences over the entire range of attribute levels, while

Figure 3: Predicted performance by the number of questions (Q, with $H = 100$)



our proposed framework aims to focus on high preference region at the aggregate level for the best prediction in high dimensions. The benchmark study is to present how our goal-directed question-selection method works differently from the existing preference elicitation methods, and to shed some light on conditions where one type of selection criterion is more appropriate than the other.

As a benchmark, we present the polyhedral method for adaptive choice-based conjoint analysis proposed by Toubia et al. (2004). The polyhedral method is one of the earliest developments in machine-learning based question-selection methods in conjoint analysis, and shown to be very efficient in reducing uncertainty in partworth parameter estimation relative to traditional conjoint analysis methods.[12] The polyhedral method iteratively selects a respondent-level choice set after a respondent completes each choice task. A choice task with n profiles generates $n\,(n-1)\,/2$ inequalities that identify an uncertainty polyhedron, where $\left(X_j - X_{j'}\right) \cdot \boldsymbol{\beta}_h \geq 0$, if a profile j is chosen by a respondent h over all other profiles j' in the choice task. The polyhedron represents a feasible region of partworth utility parameters $\boldsymbol{\beta}_h$ based on the previous choice task. The analytic center of the polyhedron estimates the partworth vector, and the distance from

[12]We believe that the polyhedral method is an appropriate benchmark, as more recent works in question-selection methods are theoretically similar in geometrically reducing uncertainty of preference parameter estimates and present empirical improvements toward the polyhedral method (e.g., Toubia et al. 2007, Sauré and Vielma 2017).

the analytic center to edge represents the uncertainty.

Several heuristic criteria were developed by Toubia et al. (2004) to select n profiles that partition the polyhedron into approximately equal parts. The heuristic criteria prioritize profiles, where corresponding inequalities cut the polyhedron in the direction that is perpendicular to long axes. Therefore, the subsequent choice updates and quickly reduces the uncertainty polyhedron with additional $n\left(n-1\right)/2$ inequalities. We replicate the question selection criteria in Toubia et al. (2004) and simulate respondents' choices in the same dimensionality of attribute levels (4 attributes with 5, 8, 11 and 12 levels each), total number of respondents (500), and number of questions per respondent (25) as in our simulation study in section 3.3.1. Partworth parameters are estimated by Hierarchical Bayesian multinomial logit (Rossi et al. 2005) consistent with Toubia et al. (2004).

Table 1 compares predictive performance of the proposed framework with 25 questions per round, 100 respondents per round, and 5 rounds in total with that of the polyhedral method with the same number of questions (25) and respondents (500). The performance measures used are utility regret for top three product profiles predicted, out-of-sample correlation between observed and predicted utilities of all $5,279$ profiles, and mean squared deviation of estimated values of parameters from true values. The predicted top three profiles by the proposed framework are (1st, 5th, 3rd) in true values, while those by the polyhedral method were (6th, 9th, 11th) in true values, leading to lower utility regret of top three profiles in the proposed method. This indicates that the proposed framework works more efficiently to navigate highly preferred profiles when experimental resources are limited relative to the dimensionality.

However, the polyhedral method outperforms the proposed framework in recovering overall preferences, presenting more accurate out-of-sample utility prediction of all candidates and mean deviation between estimated and true parameters. This is mainly because the polyhedral method aims to increase precision of preference elicitation of overall attribute-levels including low preference region, while the proposed framework focuses more on high preference region to efficiently predict the best performing profile at the market level.

Table 2 further compares predictive performances between the proposed framework and polyhedral method in the highly preferred design profiles. The correlations between true and

Table 1: Comparison between proposed and polyhedral methods

	Utility Regret of Top 3	Out-of-sample Correlation	Mean Deviation of Parameters
Proposed Framework ($Q = 25, H = 100, T = 5$)	.036	.867	.630
Polyhedral Method with HB Estimation ($Q = 25$, 500 resp.)	.214	.952	.442

predicted utilities are higher for proposed method up to the top 400 profiles (0.8% of all candidate profiles) than for the polyhedral method. However, the correlation out of top 500 profiles (1% of all candidate profiles) is higher for the polyhedral method than for the proposed method. This result offers further distinction of our method by focusing more on the subregions of highly preferred combinations of attribute levels.

Table 2: Correlations between true and predicted utilities out of top 1% of profiles

Profiles		Proposed	Polyhedral
Top 100	(0.2%)	0.638	0.263
Top 200	(0.4%)	0.635	0.460
Top 300	(0.6%)	0.688	0.575
Top 400	(0.8%)	0.697	0.650
Top 500	(1%)	0.668	0.702

The benchmark results suggest that the proposed framework efficiently allocates experimental resources when the goal is to predict the best performing profile on the market, while the polyhedral method is more efficient when the goal is to learn consumer preferences of entire range of attribute levels. In addition, the computation of adaptive criteria in the proposed framework occurs only at the end of each round (4 times in the simulation), while that in the polyhedral method occurs after each respondent completes one question ($12,500$ times $= 25$ x 500 in the simulation). Therefore, with a relevant goal to achieve, the proposed framework can be an efficient and computationally light alternative to the existing adaptive preference

elicitation methods for identifying high-performance profiles.

4 Empirical Application

The proposed framework is applied to a package design project for a consumer packaged good of a leading consumer products manufacturer.[13] The manufacturer's goal is to develop the optimal design in the presence of high-dimensional and sparse parameter space avoiding a poor combination of the best levels of main effects. The expected improvement criterion is implemented to search for the most promising design concepts with highest potential, using the parameter estimates by SSVS method. The study is conducted online with high-resolution images of hypothetical product packages.

4.1 Experimental setting

The product package consists of four design attributes including visual image of the product, claim statement of key features, name of materials used in the product, and subbrand name, as described in Table 3. The design element of the main brand name is fixed, so that is not included in the design experiment. The manufacturer selects the candidate attribute levels using domain knowledge including 12 product images, 11 claim statements, 6 materials, and 12 subbrand names. One level from each attribute is considered to be as baseline with partworth preferences of zero for identification purpose. The baseline attributes construct the common outside option in the experiment.

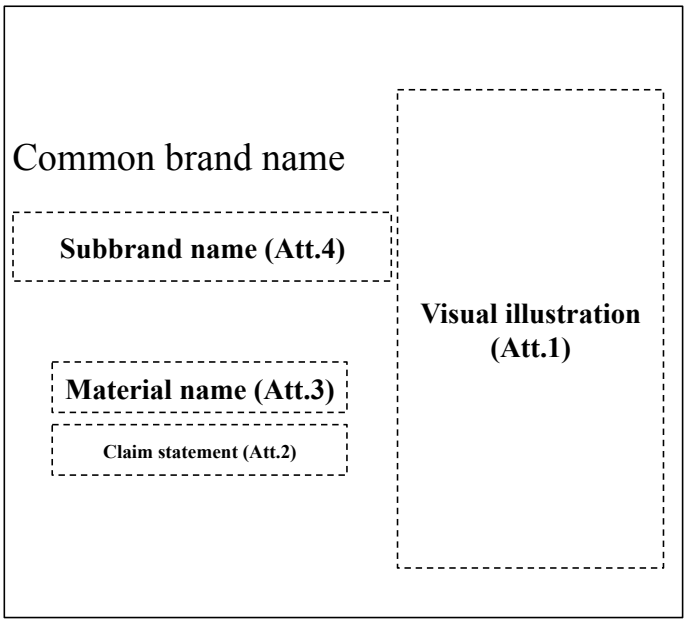
Table 3: Description of attributes

Attribute codes	Numbers of levels	Description	Visibility
Att.1	12	Visual illustration of the product	High
Att.2	11	Claim statement of the key strength	Low
Att.3	6	Name of material	Medium
Att.4	12	Subbrand name	Medium

[13]We note that the empirical application was conducted in collaboration with a nationally-known, market-leading manufacturer of consumer goods as a part of their large-scale R&D project.

Figure 4 illustrates the design elements of the product package. Attribute 1 (visual illustration) is the largest design element of the package in addition to the common main brand, so it is highly visible to respondents. Attribute 2 (claim statement) is at the bottom of the package with a small font, but includes important information to respondents. Attribute 3 (material) is placed right above the claim statement with a larger font. Attribute 4 (subbrand name) is placed right below the common brand name with a similar sized font as attribute 3. The design attributes' visibility may correlate with the size, but it does not necessarily reflect the importance of information.

Figure 4: Location of design attributes in the package



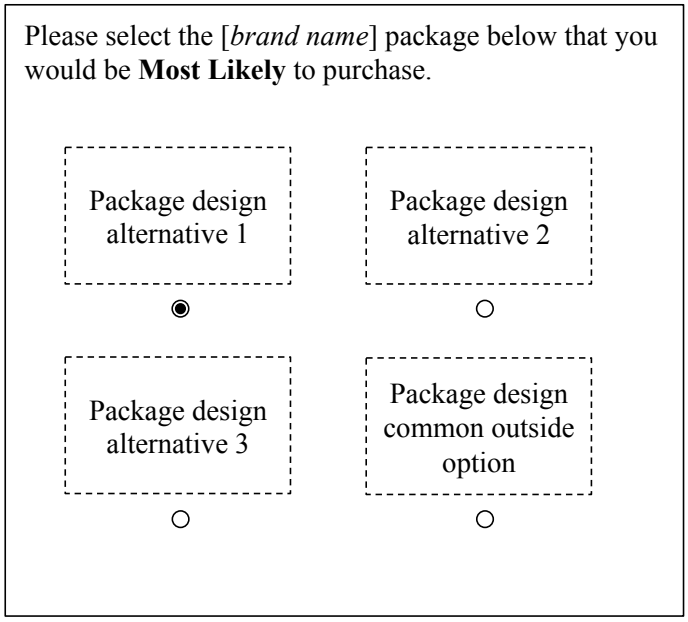
The number of rounds in the sequential experiment is predetermined as five, considering the size of data required for accurate parameter estimation and the manufacturer's typical budget limitation for R&D projects. About 450 respondents per round were participated from the U.S. and the U.K. as in Table 4. All participants confirmed that they are active users of the focal product category through screening questions. The proposed framework is an aggregate level sequential testing, so sampling with or without replacement does not affect the outcome. Simulation studies in Appendix D confirm that the true best profile was predicted and the true partworth parameters were recovered before or at the fifth round in nearly identical dimensions.

Table 4: Summary of sample sizes in each round of experiment

	Round 1	Round 2	Round 3	Round 4	Round 5
U.S.	228	223	237	217	233
U.K.	224	227	214	233	218
Total	452	450	451	450	451

Respondents receive three hypothetical design concepts and one common outside option in each question. They are requested to select their favorite design concepts out of four alternatives, as described in Figure 5. The displaying order of the four design concepts is randomized for each respondent to avoid any location effect. Respondents can enlarge the pictures of each of the given package design concepts to the full screen mode for evaluation.

Figure 5: Screen layout for the conjoint experiment



Each respondent answers 23 questions in each round. They are randomly split into five groups receiving the different product profiles for evaluation.[14] The number of alternatives tested in one round is 345 ($= 23 \times 3 \times 5$). The product profiles to be tested in the first round

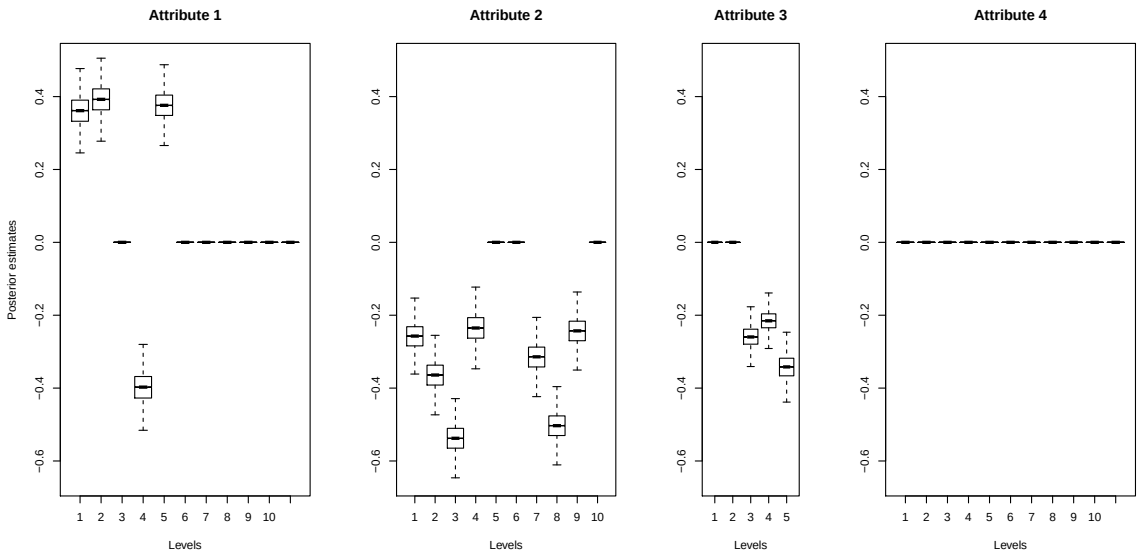
[14]The number of groups is determined based on a sensitivity test to balance between the market-share stability and the number of product profiles covered as shown in Appendix D. If there are more respondents in each group with smaller number of groups, it enhances the stability of market share calculation. On the other hand, if there are more groups with smaller number of respondents, it increases the number of questions, so more product profiles can be tested with the same number of total respondents.

are selected by classical criteria. The partworth preference parameters are estimated at the end of data collection in each round (t) using the SSVS method, and the design profiles to be tested in the next round ($t+1$) are determined by the proposed expected improvement criterion conditional on the parameter estimates. The same procedure is iterated until the end of the fifth round.

4.2 Parameter estimates

Figure 6 presents the summary of posterior estimates for main effect parameters using the data from all five rounds of experiment. Attribute 1 (visual illustration) is the most important design attribute in terms of the variation across levels, while attribute 4 (subbrand name) does not affect preferences. The visual element is the largest part in the package design, so it may attract the highest level of attention from respondents.

Figure 6: Posterior Estimates for Main-effect Partworth Parameters



Posterior estimates of each attribute level are interpreted as relative preferences to the baseline level. The levels 1, 2, and 4 in attribute 1 are preferred to the baseline level, and level 2 is the most preferred visual illustration among all candidates. The preferences of levels 5, 6, and 10 in attribute 2 are the same level as the baseline. So, the four levels in attribute 2 including the baseline are equally preferred claim statements. Similarly, the levels 1 and 2 in attribute

3 are equally preferred materials to the baseline level. Subbrand names (attribute 4) do not differentiate preferences of package designs, except for level 7 slightly less preferred to all other levels in terms of posterior means.

Among 474 potential interaction parameters, the number of non-zero interaction effects in terms of posterior means is 62 and two interaction effects out of them are significantly different from zero at the 95% credible intervals. The effect of interaction parameters is discussed in the subsequent sections.

4.3 Optimal design

The optimal package design is determined by the expected utilities of candidate product profiles using the posterior distributions of the partworth parameter estimates $\hat{\boldsymbol{\beta}}$ after the five rounds of the experiment:

$$j^* = \underset{j \in J}{\operatorname{argmax}} \, \mathbb{E}\{u\big(X_j; \hat{\boldsymbol{\beta}}\big)\} \tag{6}$$

where $u\,(\cdot)$ is the posterior distribution of respondents' preference defined in Equation (1). This final outcome returns the highest expected utility among all candidates, so it is expected to achieve the highest market share. As it includes both main and interaction effects parameters, the optimal design is selected considering the synergistic effects among attributes.

Table 5 presents the attribute levels of the optimal design concept comparing to the individual attribute with the highest preferences. Many attribute levels are indistinguishable in terms of partworth preferences, except for attribute 1. The best design concept is, therefore, finally determined by the interaction effects. The interaction effects do not only affect the prediction, but also provide inferences in preferred or less-preferred combinations. The 62 non-zero interaction parameter estimates are reported in Appendix E.

5 Discussion

The empirical results indicate that the optimal design is affected by the presence of interactive effects among the attributes. Two natural questions that arise are i) whether the proposed

Table 5: Attribute levels in the best predicted alternative

	Attribute levels in the best alternative	Attribute levels with the highest partworths
Att.1	2	2
Att.2	0	0 and 6
Att.3	1	0, 1, and 2
Att.4	1	Ties except for 7

criterion leads to the evaluation of design concepts with more appropriate interactions, and ii) whether the optimal design by the proposed framework is actually preferred to designs suggested by other methods.

The subsequent subsections first discuss the effects of selection criteria on the frequencies of appropriate interactions evaluated in the experiment. Then, it presents a validation task to evaluate the performance of the proposed framework relative to other popular methods in practice. For validation, the collaborating firm implemented three more R&D experiments using competing methods, and we conducted another separate survey to evaluate three best profiles suggested by the competing methods against the one by our proposed framework.

5.1 Evaluation frequency of interaction effects

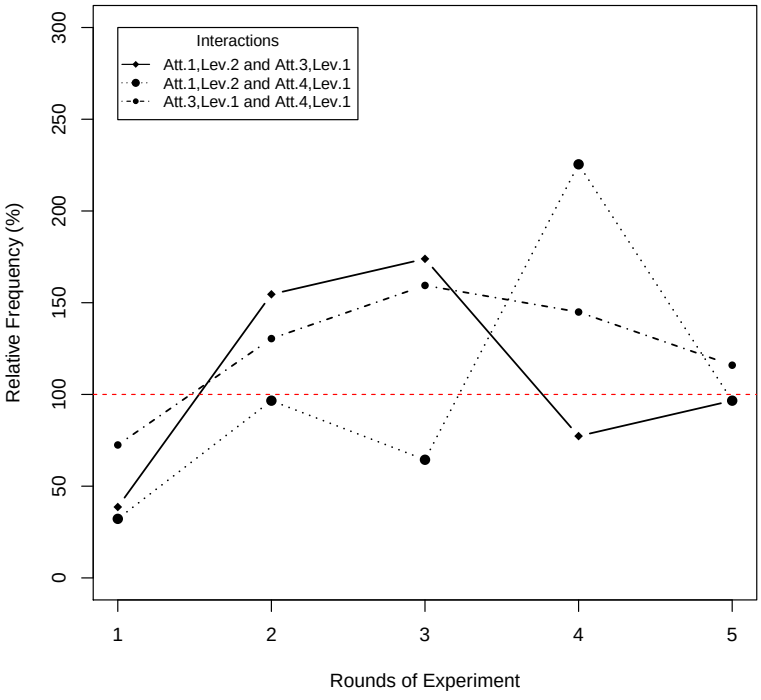
The proposed expected improvement criterion is designed to focus on the *combinations* of attributes with high potential. The key is to ensure the inclusion of the appropriate interaction effects, as not all potential interactions can be evaluated. If the proposed criteria worked as designed, the interaction effects appearing in the optimal design should have been evaluated more frequently than the expected frequency out of random draws. The relative frequency of evaluation for an interaction effect between level k of attribute a and level l of attribute b is given by:

$$\text{RelFreq}_t\left(a_k, b_l\right) = \frac{\text{ObsFreq}_t\left(a_k, b_l\right)}{\text{ExpFreq}\left(a_k, b_l\right)} \times 100$$

where ObsFreq_t is the number of product profiles including both a_k and b_l evaluated in round t, and ExpFreq is the expected number of appearances of the combination if they are randomly drawn proportionally to the total frequency in all potential candidates.

Figure 7 presents the observed relative frequencies of key interaction effects appearing in the product profiles that are evaluated in the five rounds of experiment. The interaction effects listed are those that appeared in the most preferred design concept. The dotted line indicates the expected frequency of evaluation, if candidates are randomly selected without applying proposed adaptive selection criteria. If the key interaction effects are more frequently selected to evaluate in the earlier rounds, it is highly likely to find the best design concept without too many rounds of iterations.

Figure 7: Relative observed frequency of important interaction effects



The interaction between level 1 of attribute 3 and level 1 of attribute 4 are evaluated 38% more frequently than expected in rounds two to five, where the proposed criterion is applied. The other two interactions (level 2 of attribute 1/ level 1 of attribute 3, and level 2 of attribute

1/ level 1 of attribute 4) are evaluated 26% and 21% more frequently than expected in rounds two to five. The three interaction effects are evaluated less frequently than expected in round one where the proposed criterion is not applied. Once the expected improvement criterion is applied at the end of round one, they are commonly included with higher frequency in round two of the experiment. The evaluation frequency is decreased in the later rounds, as product profiles in the high preference region including those interaction effects are mostly evaluated in the earlier rounds.

The relative frequency illustrates that the proposed criterion allocates the limited number of questions in a more efficient way to search for the potentially best combinations of design concepts. The respondents are directly exposed to the combinations with appropriate interaction effects. This confirms that the important interactive effects between attributes are less likely to be omitted in the prediction of optimal design, as respondents make head-to-head comparisons among highly preferred combinations.

5.2 Validation of the optimal design

The empirical application in section 4 was conducted as part of a large-scale R&D project by the manufacturer, which consists of four separate product design experiments including our proposed framework. The manufacturer has relied for a long time on methodologies offered by commercial vendors, Nielson's optimizer with evolutionary genetic algorithm and Sawtooth Software's choice based conjoint (CBC) experiment.[15] They also adopted and developed a machine-learning based query-selection method, called optimal Bayesian recommendation set (Viappiani and Boutilier 2010). Therefore, the R&D project produced four different product profiles recommended by each method - the proposed one, genetic algorithm, standard CBC, and Bayesian recommendation set.

In addition, an additional validation survey was conducted to directly compare the proposed optimal profile with the other three profiles created by the competing methods. A new set of individuals was selected and responded to one choice task of their favorite design among the

[15] Nielson's optimizer and Sawtooth CBC are two of the most popular methods for product design in practice. Their clients include nationally-known manufacturers in various industries, such as Unilever and Johnson & Johnson.

four profiles and a no-choice option.

5.2.1 Description of benchmark methods

The three other design experiments were conducted for the exactly same product package described in section 4. They were implemented under supervision of the collaborating firm with software providers, and we have limited information on details of implementation except for the final outcome. Therefore, we briefly describe the three benchmark methods at a conceptual level.

Nielson's genetic algorithm. Nielson's optimizer adaptively searches for the best product designs at the individual level using interactive genetic algorithm based on Malek (2001). The genetic algorithm is a heuristic approach to mimic nature's evolutionary process, where superior ones eventually survive (Balakrishnan and Jacob 1996). The questionnaire starts with a random initial set of product profiles. Subsequent sets of questions present superior offspring of product profiles in the previous round, i.e., combination of preferred attribute levels. The algorithm allows mutation in general for exploration purposes. Empirical studies have shown that the outcome of genetic algorithm is often close to optimal and outperformed other existing heuristic methods (e.g., Balakrishnan and Jacob 1996), as multiple iterations of evolution improve the fitment of the outcome.

Sawtooth CBC. The choice-based conjoint (CBC) method is a standard hierarchical Bayes conjoint model provided by Sawtooth software. They offer 30 different sets of 20 predetermined choice-tasks each, using several randomized design criteria, such as orthogonal design. Respondents first build their own designs using a graphical configurator, then the algorithm determines which questionnaire set the respondents receive out of 30 sets. The responses out of 20 questions per individual are analyzed by hierarchical Bayesian method, so the partworth preference parameters are estimated at the individual level accounting for heterogeneity.

Bayesian recommendation set. A machine-learning algorithm searching for optimal recommendation sets (Viappiani and Boutilier 2010) is adopted by the researchers in the manufacturer. Viappiani and Boutilier (2010) show that the myopically optimal choice set in an adaptive experiment is equivalent to the optimal recommendation set of the same size, i.e., a set

of product profiles that maximizes the respondent's expected utility. In the sequential process, it presents a set of product profiles to test in the next round that maximizes their sum of expected utilities using partworth preference parameter estimates in the previous round. The part-worth parameters are estimated by hierarchical Bayesian method accounting for heterogeneity of individual respondents.

5.2.2 Validation survey

The four separate experiments including the proposed method and three benchmark methods described in section 5.2.1 result in four different optimal design profiles for the identical product package. All four methods predicted four different best profiles with internal validity according to preference estimates from each model, but separate results are not able to present external validity. Therefore, we conduct a separate validation survey to compare four different design profiles out of different experimental methods.

Participants include 523 individuals from the U.S. (= 266) and the U.K. (= 257) and are active users of the focal product category. All respondents receive one question of choosing their favorite design concept out of four product profiles including the proposed one and no choice option. The orders of four design concepts are randomized to avoid location effects.

Table 6: Validation experiment

	Choice frequency	Proportion	Relative share lift by the proposed design
Proposed method	140	26.8%	
Genetic algorithm	130	24.9%	8%
Standard CBC	123	23.5%	14%
Optimal recommendation set	92	17.6%	52%
No-choice	38	7.3%	
Total	523	100.0%	

Table 6 presents the *observed* shares of the four design concepts generated by different methods in the manufacturer's R&D project. The optimal design created by the proposed framework is the most preferred design concept out of the four product profiles, each of which

is predicted as the best design by different methods. The proposed design lifts the observed share by 8% relative to the design by genetic algorithm, and by 14% and 52% relative to the standard choice based conjoint and the machine learning method, respectively. We note that all three benchmark methods fully controlled respondents' heterogeneity, and especially genetic algorithm is provided at a very high cost to the manufacturer. Though we are not able to offer the market share prediction based on this result, the observed improvement is potentially significant considering that the manufacturer's revenue per brand is over $1 billion on average.

The validation results show that the proposed framework identifies the optimal product profile by prioritizing appropriate combinations of attributes in the sequential test. The standard choice-based conjoint analysis relies on a classical experimental design, which frequently produces a main-effect design without interactions. Genetic algorithm (Malek 2001) and optimal Bayesian recommendation set (Viappiani and Boutilier 2010) are designed to overcome such problems, but they are sensitive to the initial seed with limited exploration and rely on heuristic comparison in a subset of product profiles. The results confirm that the share model used in our proposed framework is suitable for identifying market-share maximizing design concepts.

6 Conclusion

This paper proposes a new approach to optimal product design in high dimensions using sequential experiments. Product profiles are prioritized for inclusion if they can improve on the outcome of the current best design. The expected improvement criterion is operationalized by an integration of upper tail in the posterior distribution of aggregate market share. A stochastic search variable selection method reduces the dimensionality of the model by selecting relevant variables. We demonstrate that the proposed framework identifies the best design in a large scale R&D project conducted by a major packaged goods company.

The proposed criterion of expected improvement integrates preference elicitation and goal maximization and systematically balances between exploitation of high performance regions and exploration of high potential regions leading to efficient search for product profiles to test. Its evaluation in aggregate market shares directly operationalizes a firm's goal function in prod-

uct design problems. The empirical result shows that the proposed method using an aggregate market share model produces a better design concept than those suggested by competing methods, including a heterogeneous hierarchical Bayes conjoint model and commercialized methods controlling for heterogeneity. It also presents that the proposed method conceptualizing one-step-ahead optimality at the whole product level is less likely to omit important interactive effects.

The prioritization of candidates with thicker or wider upper tails in the proposed criterion is consistent with the managerial goal of marketing practice. Marketing managers often need to focus their attention on the extremes of distributions on the market (Allenby and Ginter 1995) than the average, or expected outcomes. That is, the experimental goal function should favor a product that is the most likely to improve the current status, and this is often not compatible with wanting to learn about all the part-worths associated with the decision model as implicitly assumed with using standard design criteria such as D-optimality. Our proposed framework provides an experimental criterion that is more consistent with the objective of design projects in practice.

Our modeling framework can be applied to many high-dimensional design settings, such as identifying brand logos, optimal advertising campaigns, etc. It can also be applied to other R&D projects with horizontal variation in the attribute levels. The proposed framework is especially effective when the design attributes contain a large number of levels, and the evaluation of all potential candidates is infeasible.

Appendix A: The proof of the equivalence of rank orders between upper tail integration and expected return to search

The rank orders of expected return to search is preserved in the values of upper-tail integration. The proof is shown in two separate cases of different rank orders in lower-tail values.

Case 1: $\int_{-\infty}^{z_t} z_t dF\left(u; X_j, \boldsymbol{\beta}\right) \geq \int_{-\infty}^{z_t} z_t dF\left(u; X_{j'}, \boldsymbol{\beta}\right)$ (The lower-tail integration value of j is higher than j'.)

If $V'\left(X_j\right) > V'\left(X_{j'}\right)$, which is equivalent to

$$\int_{z_t}^{\infty} u\left(X_j; \boldsymbol{\beta}\right) dF\left(u; X_j, \boldsymbol{\beta}\right) > \int_{z_t}^{\infty} u\left(X_{j'}; \boldsymbol{\beta}\right) dF\left(u; X_{j'}, \boldsymbol{\beta}\right),$$

then it is straightforward to show

$$\int_{-\infty}^{z_t} z_t dF\left(u; X_j, \boldsymbol{\beta}\right) + \int_{z_t}^{\infty} u\left(X_j; \boldsymbol{\beta}\right) dF\left(u; X_j, \boldsymbol{\beta}\right) >$$
$$\int_{-\infty}^{z_t} z_t dF\left(u; X_{j'}, \boldsymbol{\beta}\right) + \int_{z_t}^{\infty} u\left(X_{j'}; \boldsymbol{\beta}\right) dF\left(u; X_{j'}, \boldsymbol{\beta}\right), \forall j \neq j'$$

Therefore, if $V'\left(X_j\right) > V'\left(X_{j'}\right)$, then $V\left(X_j\right) > V\left(X_{j'}\right)$, when $\int_{-\infty}^{z_t} z_t dF\left(u; X_j, \boldsymbol{\beta}\right) \geq \int_{-\infty}^{z_t} z_t dF\left(u; X_{j'}, \boldsymbol{\beta}\right)$.

Case 2: $\int_{-\infty}^{z_t} z_t dF\left(u; X_j, \boldsymbol{\beta}\right) < \int_{-\infty}^{z_t} z_t dF\left(u; X_{j'}, \boldsymbol{\beta}\right)$ (The lower tail integration value of j' is higher than j.)

It is straightforward to show

$$\int_{-\infty}^{z_t} dF\left(u; X_j, \boldsymbol{\beta}\right) < \int_{-\infty}^{z_t} dF\left(u; X_{j'}, \boldsymbol{\beta}\right),$$

because z_t is a constant. The following relationship is obtained by rearranging the previous inequality:

$$\int_{-\infty}^{z_t} \left\{dF\left(u; X_{j'}, \boldsymbol{\beta}\right) - dF\left(u; X_j, \boldsymbol{\beta}\right)\right\} = \int_{z_t}^{\infty} \left\{dF\left(u; X_j, \boldsymbol{\beta}\right) - dF\left(u; X_{j'}, \boldsymbol{\beta}\right)\right\} > 0$$

If $V'(X_j) > V'(X_{j'})$, which is equivalent to

$$\int_{z_t}^{\infty} u(X_j;\boldsymbol{\beta})\, dF(u;X_j,\boldsymbol{\beta}) > \int_{z_t}^{\infty} u(X_{j'};\boldsymbol{\beta})\, dF(u;X_{j'},\boldsymbol{\beta}),$$

then the following inequality holds by the relationships above:

$$V(X_j) - V(X_{j'})$$

$$= \left\{ \int_{-\infty}^{z_t} z_t dF(u;X_j,\boldsymbol{\beta}) + \int_{z_t}^{\infty} u(X_j;\boldsymbol{\beta})\, dF(u;X_j,\boldsymbol{\beta}) \right\} - \left\{ \int_{-\infty}^{z_t} z_t dF(u;X_{j'},\boldsymbol{\beta}) + \int_{z_t}^{\infty} u(X_{j'};\boldsymbol{\beta})\, dF(u;X_{j'},\boldsymbol{\beta}) \right\}$$

$$= \int_{-\infty}^{z_t} z_t \left\{ dF(u;X_j,\boldsymbol{\beta}) - dF(u;X_{j'},\boldsymbol{\beta}) \right\} + \int_{z_t}^{\infty} u(X_j;\boldsymbol{\beta}) \left\{ dF(u;X_j,\boldsymbol{\beta}) - dF(u;X_{j'},\boldsymbol{\beta}) \right\}$$

$$= \int_{z_t}^{\infty} -z_t \left\{ dF(u;X_j,\boldsymbol{\beta}) - dF(u;X_{j'},\boldsymbol{\beta}) \right\} + \int_{z_t}^{\infty} u(X_j;\boldsymbol{\beta}) \left\{ dF(u;X_j,\boldsymbol{\beta}) - dF(u;X_{j'},\boldsymbol{\beta}) \right\}$$

$$= \int_{z_t}^{\infty} \{u(X_j;\boldsymbol{\beta}) - z_t\} \left\{ dF(u;X_j,\boldsymbol{\beta}) - dF(u;X_{j'},\boldsymbol{\beta}) \right\} > 0$$

This is because

$$u(X_j;\boldsymbol{\beta}) - z_t > 0,\ u \in [z_t, \infty)$$

and

$$\int_{z_t}^{\infty} \left\{ dF(u;X_j,\boldsymbol{\beta}) - dF(u;X_{j'},\boldsymbol{\beta}) \right\} > 0$$

Therefore, if $V'(X_j) > V'(X_{j'})$, then $V(X_j) > V(X_{j'})$, when $\int_{-\infty}^{z_t} z_t dF(u;X_j,\boldsymbol{\beta}) < \int_{-\infty}^{z_t} z_t dF(u;X_{j'},\boldsymbol{\beta})$.

The cases 1 and 2 cover all potential regions of relationships in lower tail integration values. Combining both cases, it is shown that if $V'(X_j) > V'(X_{j'})$, then $V(X_j) > V(X_{j'})$, $\forall j \neq j'$.

Appendix B: Stochastic search variable selection method

This appendix provides the estimation procedure for the models in the empirical analysis. We follow the Gibbs sampling method proposed by George and McCulloch (1993). The details of derivation of posterior distributions are described in George and McCulloch (1993, 1997). The following three steps summarize the procedure to draw from the posterior distributions.

First, the partworth parameters are drawn from the multivariate normal distribution conditional on the variable selection γ, variance σ^2, and the data. If a variable is irrelevant, it is drawn from a mass concentrated around zero.

$$[\boldsymbol{\beta}|\gamma, \sigma^2, \boldsymbol{X}, \boldsymbol{Y}] = N\{[\sigma^{-2}\boldsymbol{X}'\boldsymbol{X} + (\boldsymbol{D}_\gamma \boldsymbol{R} \boldsymbol{D}_\gamma)^{-1}]^{-1}\sigma^{-2}\boldsymbol{X}'\boldsymbol{Y}, \sigma^{-2}\boldsymbol{X}'\boldsymbol{X} + (\boldsymbol{D}_\gamma \boldsymbol{R} \boldsymbol{D}_\gamma)^{-1}]^{-1}\}$$

Second, the variance is drawn from the inverted gamma distribution conditional on $\boldsymbol{\beta}$ and the data. The number of observations is nQt in round t as the estimation is on the cumulative data up to the current round of the experiment.

$$[\sigma^2|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}] = IG\left(\frac{nQt + \kappa}{2}, \frac{|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|^2 + \kappa\psi}{2}\right)$$

Third, the variable selection index of an individual variable γ_i is drawn from Bernoulli distribution with probability of the fraction of conditional likelihoods. The selection of other variables are considered to be as given at each draw of γ_i.

$$[\gamma_i|\boldsymbol{\beta}, \gamma_{-i}] = Bernoulli\,(p_i)\,,\; p_i = \frac{[\boldsymbol{\beta}|\gamma_i = 1, \boldsymbol{X}, \boldsymbol{Y}]}{[\boldsymbol{\beta}|\gamma_i = 1, \boldsymbol{X}, \boldsymbol{Y}] + [\boldsymbol{\beta}|\gamma_i = 0, \boldsymbol{X}, \boldsymbol{Y}]}$$

The three steps are iterated until convergence.

Appendix C: Benchmark study of estimation methods

A simulation study is used to evaluate how well the SSVS estimation works along with the proposed question-selection criterion. We evaluate the accuracy of parameter estimates for both main and interaction effects using SSVS (George and McCulloch 1997), Lasso (Tibshirani 1996), and standard Bayesian regression (Rossi et al. 2005).

We simulate choice share data from the true partworth preference parameters and assume that the number of product profiles for testing in each question is $4\,(= n)$ and outside option. It is also assumed that there are 20 questions per round, so each round of the experiment evaluates $80\,(= nQ)$ product profiles in addition to the outside option. The number of rounds is assumed to be $5\,(= T)$, for a total of $400\,(= nQT)$ product profile evaluations plus the outside option. Each dependent variable $(\ln S_{jq} - \ln S_{0q})$ for a design concept j in question q is simulated from Equation (1) with true values of $\boldsymbol{\beta}$ and random draws of product-specific error terms, ξ_{jq}, from a normal distribution with a mean of zero and a standard deviation of one.

The product in the simulated experiment has four attributes with $(5, 8, 11, 12)$ levels each. The true partworth preference parameters are designed in a way that the true best design concept is not a combination of the most preferred individual attribute-levels, such that interactions among attributes are large and affect the most preferred alternative. The partworth utility of common outside option across questions is normalized to be zero. The dummy coding of all design attributes leads to 32 main effects and 369 two-way interaction effects, and there are $5,279$ candidate product profiles in total.

Table 7 compares the three different estimation methods using the mean absolute percentage errors (MAPE) between observed and predicted values of dependent variables.[16] The SSVS estimation presents the lowest MAPE, indicating that mean preference estimates were more accurately estimated with SSVS than with other methods.

Table 7: Out-of-sample predictive fits in mean absolute percentage errors

	MAPE
SSVS	0.156
Lasso	0.169
Bayes Regression	0.218

Posterior estimates also suggest that the SSVS method works well in estimation of the main and interaction effects and in model selection. Figure 8 presents the main-effect posterior

[16]For predictive fit measurement, a percentage measure is used to incorporate the potential variation in the observed values of the dependent variable. It prevents to overweight a small error in predicting a large observed value.

estimates for attribute 2 using the three estimation methods. The vertical lines stand for the 95% credible intervals, and the dotted lines are the 45-degree lines, where the true values and estimates are identical. It shows that the true values of all levels of main effect estimates for the attribute are accurately recovered by SSVS and Lasso preserving the rank order. SSVS presents tighter credible intervals than Lasso. The standard Bayesian regression failed to recover the rank order of levels, as interaction effects are not well identified from the main effects given small sample size. The credible intervals in Bayesian regression are much wider than SSVS and Lasso, due to the dimensionality. This tendency is similar in other attributes.

Figure 8: Posterior Estimates for Attribute 2

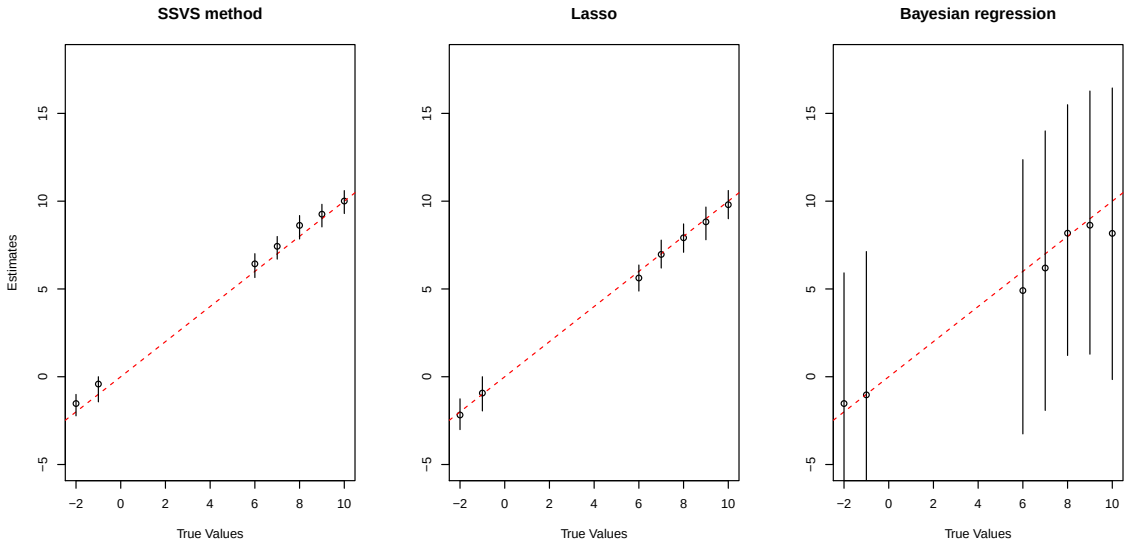
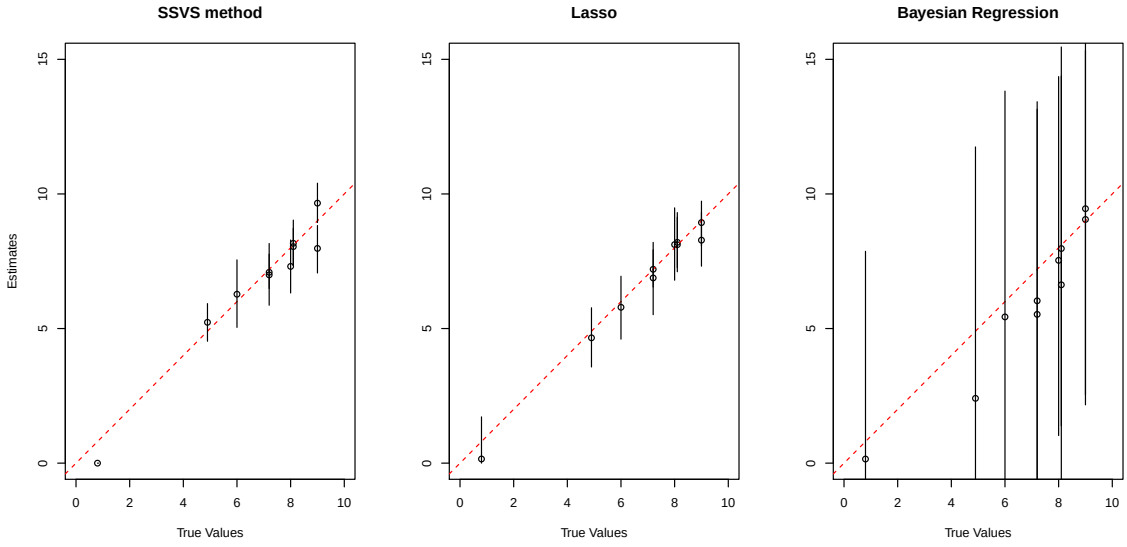


Figure 9 presents the interaction effect posterior estimates using the three estimation methods. Both SSVS and Lasso are reasonably accurate in recovering true values of non-zero interaction parameter estimates. Similar to the main effect estimates, the SSVS presents tighter credible intervals than Lasso, and the standard Bayesian regression presents very wide credible intervals.

The SSVS estimation presents higher accuracy in model selection as shown in Table 8. The SSVS method accurately predicted 99.44% of irrelevant interaction effects successfully reducing the 'dimensionality' of non-zero coefficients, while Lasso and standard Bayesian regression only

Figure 9: Posterior Estimates for Interaction Effects



predicted 27.3% and 1.67%, respectively.[17] The non-zero interaction effects are detected as relevant variables in all estimation methods. The SSVS method estimated one non-zero interaction parameter as zero, but this is because the true value is relatively close to zero.

Table 8: Prediction accuracy of interaction effects

	SSVS		Lasso		Bayes.regression	
	True	False	True	False	True	False
Non-zero Interactions (10 in total)	90.00%	10.00%	100.00%	0.00%	100.00%	0.00%
Zero Interactions (359 in total)	99.44%	0.56%	27.30%	72.70%	1.67%	98.33%

Note 1: Estimates less than 0.01 are considered to be zero.
Note 2: True value of the interaction parameter that is mispredicted by SSVS is 0.8.

Appendix D: Model sensitivity test to the number of groups, the total number of rounds, and respondents' heterogeneity

This appendix presents a simulation study to test how sensitive the proposed framework is to the variation in the number of respondents per question and the total number of rounds of a sequential experiment in the presence of heterogeneity. We assume that there are 450

[17] The threshold to be *zero* is 0.01 in Table 8. SSVS results were robust using various values of the threshold (e.g., 0.5, 0.1, 0.05, 0.01, and 0.001), while other methods became worse with smaller values.

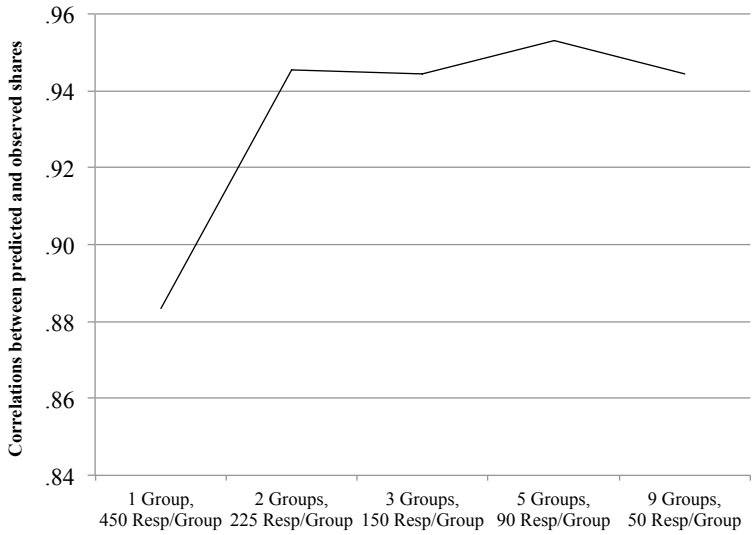
heterogeneous respondents, and test sensitivity in predictive performance when respondents are split into 1, 2, 3, 5, and 9 groups. For example, if we split respondents into 5 groups, each group includes 90 respondents receiving the same questions to calculate market shares, and different groups receive different sets of questions.[18]

We assume that the true data generating process is

$$S_{jq} = \int \frac{\exp\left(u\left(X_j;\boldsymbol{\beta}_h\right)\right)}{1+\sum_{j'\in J_q}\exp\left(u\left(X_{j'};\boldsymbol{\beta}_h\right)\right)} dF\left(\boldsymbol{\beta}\right)$$

where $\boldsymbol{\beta}_h \sim N\left(\bar{\boldsymbol{\beta}}, \Sigma\right)$. The market-share observations in the simulation study are generated by counting heterogeneous individuals' choices using their partworth parameters $\boldsymbol{\beta}_h$ randomly drawn from the normal distribution.[19] Each choice outcome of each individual is simulated by the true values of $\boldsymbol{\beta}_h$ and random draws from a Type I extreme value distribution with a location parameter of zero and a scale parameter of one. We estimate the aggregate-level model in Equation (1) to estimate the average preferences and predict the market shares generated from the individual-level choice model.

Figure 10: Predictive performance sensitivities by respondents grouping



[18] At a glance, one might think that splitting into a larger number of groups may be better as it increases the number of product profiles to be evaluated. However, it comes at a cost of a small number of respondents in each group, where the market share may not be representative to proceed the adaptive process. Therefore, we conduct a sensitivity test of this trade-off to determine the number of groups in our empirical application.

[19] The diagonal elements of Σ are set to be one, and off-diagonal elements are set to be zero.

Figure 11: Predictive performance sensitivities by the number of rounds

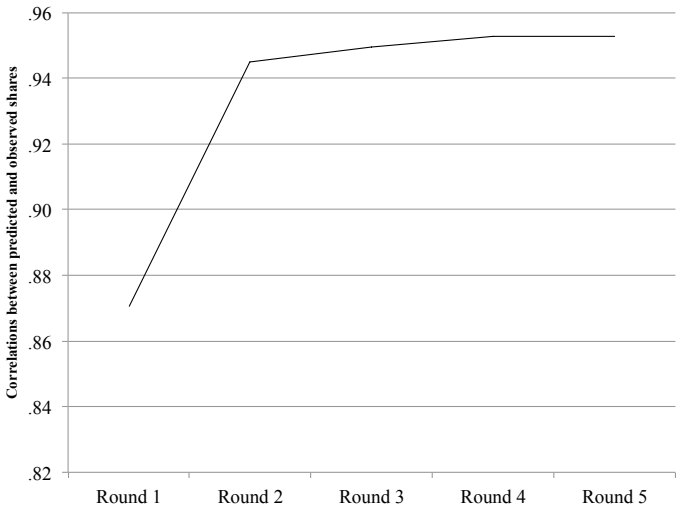
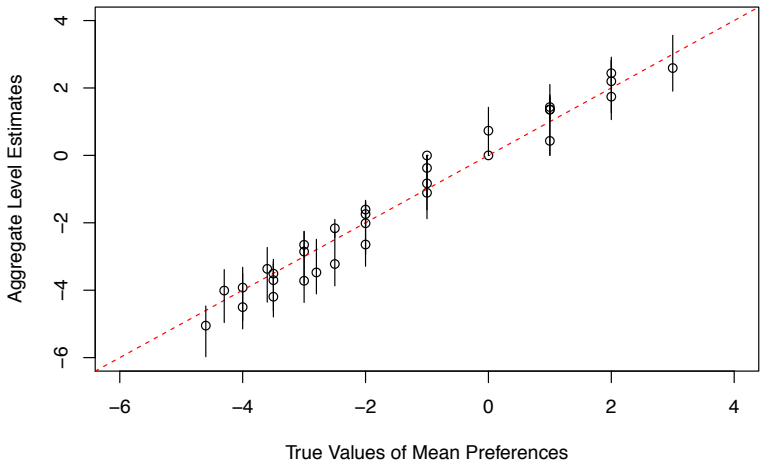


Figure 12: Parameter recovery in the aggregate-level model with 5 groups



Under the five different grouping decisions, the sequential experiment is simulated using the proposed criteria and SSVS method. Figure 10 presents the correlations between observed market shares generated by the individual-level preference parameters and predicted market shares by the aggregate-level model after the five rounds of adaptive experiments. The results indicate that splitting respondents into 5 groups with 90 respondents each shows the best predictive performance. Also, the aggregate-level model predicts the market shares generated by the individual-level preference parameters very accurately with correlation of .969. Figure 11

presents how the correlation between observed and predicted market shares change every round under the optimal group structure. The predicted market shares are already very close to the true values in round 2, and converged at round 4. This confirms that five rounds of experiment are enough for optimal product design in the given dimensionality.

The aggregate-level model not only predicts the true best design simulated by the individual-level data generating process with heterogeneity, but also recovers the true values of mean preferences in the individual-level model with accurate rank orders using only a subset of data as presented in Figure 12.

Appendix E: Interaction effects in the empirical application

Figure 13 presents a boxplot of 14 selected interaction parameter estimates, where posterior means are greater than .0001 out of 474 potential interaction effects. There are two notable negative interaction effects between level 1 of attribute 1 and level 4 of attribute 2 (12), and between level 1 of attribute 1 and level 5 of attribute 2 (13), which indicate that combinations of those attribute levels are not perceived as good design concepts as a whole. Table 9 reports all non-zero interaction parameter estimates in the empirical application, where positive posterior means served as tie-breakers of main effects.

Figure 13: Posterior Estimates for Selected Interaction Effects

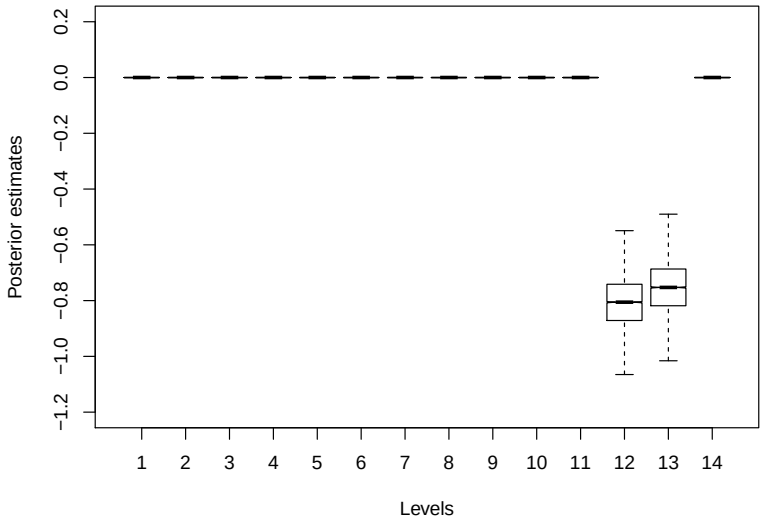


Table 9: Non-zero interaction parameter posterior estimates

1st Attributes	2nd Attributes	Mean	Std.err.	1st Attributes	2nd Attributes	Mean	Std.err.
Att.1,Lev.1	Att.2,Lev.1	.00000	(.00021)	Att.1,Lev.7	Att.2,Lev.2	.00005	(.00481)
Att.1,Lev.1	Att.2,Lev.4	-.80729	(.09983)	Att.1,Lev.7	Att.2,Lev.6	.00005	(.00503)
Att.1,Lev.1	Att.2,Lev.5	-.75285	(.09686)	Att.1,Lev.7	Att.2,Lev.8	.00782	(.06728)
Att.1,Lev.1	Att.2,Lev.6	.00000	(.00039)	Att.1,Lev.7	Att.3,Lev.2	.00035	(.01055)
Att.1,Lev.1	Att.3,Lev.3	-.00010	(.00582)	Att.1,Lev.7	Att.3,Lev.3	.00003	(.00264)
Att.1,Lev.1	Att.3,Lev.5	.00012	(.00635)	Att.1,Lev.7	Att.3,Lev.5	.00002	(.00228)
Att.1,Lev.1	Att.4,Lev.6	-.00001	(.00128)	Att.1,Lev.7	Att.4,Lev.1	.00003	(.00314)
Att.1,Lev.1	Att.4,Lev.8	.00004	(.00381)	Att.1,Lev.7	Att.4,Lev.10	.00015	(.01457)
Att.1,Lev.10	Att.3,Lev.3	.00002	(.00143)	Att.1,Lev.7	Att.4,Lev.3	.00002	(.00175)
Att.1,Lev.10	Att.4,Lev.1	.00001	(.00096)	Att.1,Lev.8	Att.2,Lev.4	.00001	(.00136)
Att.1,Lev.10	Att.4,Lev.2	-.00002	(.00233)	Att.1,Lev.8	Att.3,Lev.2	.00000	(.00021)
Att.1,Lev.10	Att.4,Lev.8	-.00003	(.00283)	Att.1,Lev.8	Att.3,Lev.5	-.00006	(.00414)
Att.1,Lev.11	Att.2,Lev.10	.00005	(.00461)	Att.1,Lev.8	Att.4,Lev.11	-.00008	(.00610)
Att.1,Lev.11	Att.2,Lev.7	.00002	(.00216)	Att.1,Lev.8	Att.4,Lev.8	.00002	(.00146)
Att.1,Lev.11	Att.3,Lev.2	.00002	(.00147)	Att.1,Lev.9	Att.2,Lev.8	.00003	(.00313)
Att.1,Lev.11	Att.3,Lev.5	.00003	(.00283)	Att.2,Lev.1	Att.3,Lev.3	.00004	(.00360)
Att.1,Lev.11	Att.4,Lev.5	.00011	(.00623)	Att.2,Lev.1	Att.4,Lev.10	.00002	(.00173)
Att.1,Lev.11	Att.4,Lev.9	.00003	(.00244)	Att.2,Lev.1	Att.4,Lev.5	.00003	(.00277)
Att.1,Lev.2	Att.2,Lev.5	-.00002	(.00176)	Att.2,Lev.10	Att.4,Lev.8	.00000	(.00030)
Att.1,Lev.2	Att.3,Lev.2	-.00003	(.00292)	Att.2,Lev.3	Att.3,Lev.1	.00000	(.00012)
Att.1,Lev.2	Att.4,Lev.1	.00003	(.00310)	Att.2,Lev.3	Att.4,Lev.7	.00002	(.00145)
Att.1,Lev.3	Att.2,Lev.2	.00000	(.00026)	Att.2,Lev.4	Att.3,Lev.3	-.00043	(.01295)
Att.1,Lev.3	Att.3,Lev.4	.00004	(.00286)	Att.2,Lev.4	Att.4,Lev.3	-.00006	(.00561)
Att.1,Lev.4	Att.2,Lev.4	.00004	(.00391)	Att.2,Lev.6	Att.3,Lev.1	.00000	(.00047)
Att.1,Lev.4	Att.4,Lev.10	.00002	(.00230)	Att.2,Lev.6	Att.4,Lev.2	-.00003	(.00251)
Att.1,Lev.6	Att.2,Lev.2	.00012	(.00674)	Att.2,Lev.6	Att.4,Lev.5	.00002	(.00174)
Att.1,Lev.6	Att.2,Lev.3	.00013	(.00750)	Att.2,Lev.7	Att.4,Lev.6	-.00003	(.00330)
Att.1,Lev.6	Att.2,Lev.9	.00005	(.00347)	Att.2,Lev.9	Att.4,Lev.1	.00002	(.00219)
Att.1,Lev.6	Att.3,Lev.3	.00005	(.00372)	Att.2,Lev.9	Att.4,Lev.7	-.00014	(.00704)
Att.1,Lev.6	Att.3,Lev.4	.00033	(.01059)	Att.3,Lev.4	Att.4,Lev.2	-.00002	(.00167)
Att.1,Lev.6	Att.3,Lev.5	-.00004	(.00372)	Att.3,Lev.4	Att.4,Lev.5	.00011	(.00567)

References

Allenby, Greg M., James L. Ginter. 1995. Using extremes to design products and segment markets. *Journal of Marketing Research* **32**(4) 392–403.

Allenby, Greg M, Peter E Rossi. 1991. There is no aggregation bias: Why macro logit models work. *Journal of Business and Economic Statistics* **9**(1) 1–14.

Balakrishnan, P. V. (Sundar), Varghese S. Jacob. 1996. Genetic algorithms for product design. *Management Science* **42**(8) 1105–1117.

Berry, Steven T. 1994. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics* **25**(2) 242–262.

Bien, Jacob, Jonathan Taylor, Robert Tibshirani. 2013. A lasso for hierarchical interactions. *The Annals of Statistics* **41**(3) 1111–1141.

Brezzi, Monica, Tze Leung Lai. 2002. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control* **27** 87–108.

Cavagnaro, Daniel R., Jay I. Myung, Mark A. Pitt, Janne V. Kujala. 2010. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation* **22** 887–905.

Dzyabura, Daria, John R. Hauser. 2011. Active machine learning for consideration heuristics. *Marketing Science* **30**(5) 801–819.

George, Edward I., Robert E. McCulloch. 1993. Variable selection via gibbs sampling. *Journal of the American Statistical Association* **88**(423) 881–889.

George, Edward I., Robert E. McCulloch. 1997. Approaches for bayesian variable selection. *Statistica Sinica* **7** 339–373.

Gilbride, Timothy J., Greg M. Allenby, Jeff D. Brazell. 2006. Models for heterogeneous variable selection. *Journal of Marketing Research* **43**(3) 420–430.

Hans, Chris. 2009. Bayesian lasso regression. *Biometrika* **96**(4) 835–845.

Huang, Dongling, Lan Luo. 2016. Consumer preference elicitation of complex products using fuzzy support vector machine active learning. *Marketing Science* **35**(3) 445–464.

Jones, Donald R., Matthias Schonlau, William J. Welch. 1998. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* **13** 455–492.

Lindley, Dennis V. 1956. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics* 986–1005.

Liu, Qing, Thomas Otter, Greg M. Allenby. 2007. Investigating endogeneity bias in marketing. *Marketing Science* **26**(5) 642–650.

Malek, Kamal M. 2001. Analytical and interpretive practices in design and new product development: Evidence from the automobile industry. Ph.D. thesis, Messachusetts Institute of Technology.

Mockus, Jonas. 1994. Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization* **4** 347–365.

Park, Trevor, George Casella. 2008. The bayesian lasso. *Journal of the American Statistical Association* **103**(482) 681–686.

Rossi, Peter E., Greg M. Allenby, Robert E. McCulloch. 2005. *Bayesian Statistics and Marketing*. John Wiley and Sons Ltd.

Sauré, Denis, Juan Pablo Vielma. 2017. Ellipsoidal methods for adaptive choice-based conjoint analysis. *Working paper* .

Schonlau, Matthias, William J. Welch, Donald R. Jones. 1997. A data-analytic approach to Bayesian global optimization. *American Statistical Association Proceedings, Section of Physical Engineering Sciences* 186–191.

Schonlau, Matthias, William J. Welch, Donald R. Jones. 1998. Global versus local search in constrained optimization of computer models. *New Developments and Applications in Experimental Design* **34** 11–25.

Schwartz, Eric M., Eric Bradlow, Peter Fader. 2016. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science* Forthcoming.

Scott, Steven L. 2010. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* **26** 639–658.

Silvey, S. 2013. *Optimal design: An introduction to the theory for parameter estimation*, vol. 1. Springer Science and Business Media.

Thompson, William R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**(285–294).

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1) 267–288.

Toubia, Olivier, John R. Hauser, Rosanna Garcia. 2007. Probabilistic polyhedral methods for adaptive choice-based conjoint analysis: Theory and application. *Marketing Science* **26**(5) 596–610.

Toubia, Olivier, John R. Houser, Duncan I. Simester. 2004. Polyhedral methods for adaptive choice-based conjoint analysis. *Journal of Marketing Research* **41**(1) 116–131.

Toubia, Olivier, Duncan I. Simester, John R. Hauser, Ely Dahan. 2003. Fast polyhedral adaptive conjoint estimation. *Marketing Science* **22**(3) 273–303.

Viappiani, Paolo, Craig Boutilier. 2010. Optimal bayesian recommendation sets and myopically optimal choice query sets. *Advances in Neural Information Processing Systems* **23** 2352–2360.

Wang, Tianhan, Craig Boutilier. 2003. Incremental utility elicitation with the minimax regret decision criterion. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. 309–316.

Weitzman, Martin L. 1979. Optimal search for the best alternative. *Econometrica* **47**(3) 641–654.