Form + Function: Optimizing Aesthetic Product Design via Adaptive, Geometrized Preference Elicitation

Namwoo Kang
Department of Mechanical Systems Engineering
Sookmyung Women's University
nwkang@sm.ac.kr

Yi Ren
Department of Mechanical Engineering
Arizona State University
yiren@asu.edu

Fred M. Feinberg
Ross School of Business and Department of Statistics
University of Michigan
feinf@umich.edu

Panos Y. Papalambros
Department of Mechanical Engineering
University of Michigan
pyp@umich.edu

November 7, 2019

The authors gratefully acknowledge the support of University of Michigan Graham Sustainability Institute Fellowship, as well as Elea Feit and Jangwon Choi for their suggestions. The first two authors contributed equally.

Under second review at *Marketing Science*; please do not quote or distribute.

Form + Function: Optimizing Aesthetic Product Design via Adaptive, Geometrized Preference Elicitation

Abstract

Visual design is critical to product success, and the subject of intensive marketing research effort. Yet visual elements, due to their holistic and interactive nature, do not lend themselves well to optimization using extant decompositional methods for preference elicitation. Here we present a systematic methodology to incorporate interactive, 3D-rendered product configurations into a conjoint-like framework. The method relies on rapid, scalable machine learning algorithms to adaptively update product designs along with standard information-oriented product attributes. At its heart is a parametric account of a product's geometry, along with a novel, adaptive "bi-level" query task that can estimate individuals' visual design form preferences and their trade-offs against such traditional elements as price and product features. We illustrate the method's performance through extensive simulations and robustness checks, a formal proof of the bi-level query methodology's domain of superiority, and a field test for the design of a mid-priced sedan, using real-time 3D rendering for an online panel. Results indicate not only substantially enhanced predictive accuracy, but two quantities beyond the reach of standard conjoint methods: trade-offs between form and function overall, and willingness-to-pay for specific design elements. Moreover – and most critically for applications – the method provides "optimal" visual designs for both individuals and model-derived or analyst-supplied consumer groupings, as well as their sensitivities to form and functional elements.

Keywords: product design optimization; conjoint analysis; machine learning; preference elicitation; visual design

1 Introduction

A product's visual form has long been acknowledged as a pivotal element of consumer choice (Kotler and Rath 1984, Bloch 1995, Veryzer and Hutchinson 1998, Dahl et al. 1999, Bloch et al. 2003, Pan et al. 2017). Firms as diverse as Bang & Olufsen, Apple, Dyson, Uniqlo, and Tesla have not only made striking visuals an emblematic linchpin of their success, but have also helped propel design, as both discipline and corporate mission, into the public sphere. Even old-line firms have taken note: IBM's product design initiative is now the world's largest, and HP relishes its rivalry with Apple in that sphere.[1]

Academic research in both marketing and engineering has consequently grappled with how to capture, and ultimately optimize, product "form," with varying definitions, terminologies, and degrees of success. Preference modeling studies incorporating visual design elements have addressed general concepts like "appearance" (Creusen and Schoormans 2005), "form" (Bloch 1995, Orsborn and Cagan 2009, Tseng et al. 2012), "styling" (Dotson et al. 2019), and "design" (Landwehr et al. 2011, Burnap et al. 2019), as well as more specific ones like "shape" (Orsborn et al. 2009, Kelly et al. 2011, Orbay et al. 2015), "silhouette" (Reid et al. 2012), and "profile" (Lai et al. 2005). "Form" – broadly construed – often assumes a central role in real-world preference modeling problems. According to Bloch (1995), form helps attract consumer attention, communicate product information, and stimulate visual pleasure, thereby generating a long-lasting perceptual impression. In practice, sales predictions can be significantly improved by accommodating form (Landwehr et al. 2011). Marketers and designers also find valuable trade-offs between form and functional attributes (Dotson et al. 2019, Sylcott et al. 2013a), and revealing such trade-offs can lead to superior balance between visual appeal and functionality (Reid et al. 2012).

Product aesthetics comprises purely visual elements like color and packaging, haptic (Peck and Childers 2003) and sensory (Krishna 2012) aspects that can be altered and optimized independently, and more nuts-and-bolts 'geometric' elements that both convey product image and constrain / interact with the internal operations of the product itself. Honing these geometrical elements is critical for efficient design of components and production processes, but mapping from the geometry of a product to how much potential consumers might like it is a complex exercise (Michalek et al. 2005, 2011). A further challenge is quantifying how elements of form preference are traded off against 'traditional' attributes like price and performance. Such trade-offs are critical

[1] IBM Design (www.ibm.com/design/teams/); Paste Magazine (www.pastemagazine.com/articles/2017/08/the-story-of-spectre-how-hp-reinvented-itself-thro.html)

in allowing designers to determine not only how important "design" is overall to a particular consumer or type of consumer, but whether specific *aspects* of design – like a low-profile car hood that may constrain the powertrain and require costly amendments to maintain crashworthiness – can be accommodated, subject to supply-side and consumer budget constraints.

Marketers have traditionally employed conjoint-based techniques to assess consumer trade-offs (Green and Srinivasan 1990; Green, Krieger, and Wind 2001), but have struggled to incorporate product form into overall preference modeling. One method of doing so involves the elicitation of adjectival descriptors. For example, the largest survey of its kind, Maritz Corp.'s New Vehicle Car Survey (NVCS), asks 250,000 new car buyers annually to rate statements like "styling is at or near the top of important characteristics in a new vehicle"; to rate the importance of, satisfaction with, and likelihood of rejecting vehicles based on interior and exterior styling; to assess trade-offs with vehicle safety; and to assess 28 "image" elements.[2] Yet this procedure still leaves designers in a quandary as to *what to actually design*: while it's helpful to learn a particular consumer likes "sleek" cars, what does such a car actually look like? Reasonable people can differ on the operationalization of adjectival labels: one consumer's sleek is another's clunky, and a sleek SUV may have a radically different visual footprint from a sleek sports car. Moreover, potential buyers (or segments thereof) can express design preferences but be in no position to enact them, due to financial, familial, or other constraints, making it difficult to know how "design," as an overall attribute – or various aspects of design, like "ruggedness" – is traded-off against more prosaic but measurably important elements like price, safety, and other features.

We later review research in preference modeling, efficient algorithms, and design optimization (Section 2); yet, to the authors' knowledge, few studies combine a *form* preference model with one for *overall* preference – that is, including "traditional", conveyable product attributes – and even these few rely on decomposition via disjoint measurement (Sylcott et al. (2013a, Dotson et al. 2019). This is reasonable for lab studies where strict demographics controls can be exacted; it is more problematic when applied to online "crowdsourced" groups that can differ in crucial characteristics. Combining or "fusing" data across such groups presents serious impediments to accurately capturing individual-level preference (Feit et al. 2010, Feit et al. 2013, Qian and Xie 2013).

[2] Specifically, the 28 image elements were: classic, responsive, youthful, bold, luxurious, prestigious, stylish, functional, safe, innovative, economical, simple, conservative, environmentally-friendly, sleek, distinctive, fun to drive, comfortable, well-engineered, rugged/tough, elegant, family-oriented, good value, sporty, powerful, sophisticated, exciting.

To overcome these challenges, we propose a new methodology that disentangles *form preference* from *overall preference* and coordinates them adaptively, in real time, while allowing respondents to manipulate 3D renderings of product geometry. The method makes use of both Bayesian methods for preference measurement and machine learning tools for rapidly adapting a multidimensional space whose *consumer-oriented visual design* characteristics (e.g., boxy, retro, etc.) do not need to be determined in advance, or indeed at any point via verbal protocols. We apply the proposed "bi-level adaptive conjoint" method to model both form preference (for vehicles) based on underlying product geometry and overall preference by revealing trade-offs between form (e.g., hood length, windshield pitch) and functional attributes (here, price and fuel efficiency). Results – via both Monte Carlo simulation and a crowdsourced experiment – suggest that the proposed method not only zeroes in on each consumer's preferred visual design, but, more generally, can both elicit superior individual-level preference estimates than conventional conjoint alone and produce "optimal" designs for analyst-supplied consumer groupings.

These benefits in turn allow, we believe for the first time, explicit measurement of trade-offs among willingness-to-pay (WTP) and design variables, e.g., whether price-sensitive consumers have marked preferences for certain styles, whether sports car buyers are less concerned about fuel economy, etc. Although the bi-level adaptive query method is novel, the set-up is 'modular' in the sense of the ability to slot in other question types, although their efficiency, scalability, and speed would need careful testing. Rather, as we emphasize throughout, the key innovation is allowing consumers to interact with a parametric geometrization of the product's topology, within a conjoint-like framework, to enable real-time, individual- or group-level design optimization.

The paper is structured as follows: Section 2 reviews related approaches from both the marketing and engineering design literatures. Section 3 proposes the "bi-level adaptive conjoint" method, whose advantages over traditional conjoint are illustrated both by extensive simulations in Section 4 and web-based panelists in Section 5. Section 6 concludes by discussing findings and potential extensions.

2 Prior Literature on Product Design Preference Optimization

The literature on product design, both as a stand-alone field and within cognate disciplines, is vast. We therefore provide concise integrative discussions of three lines of prior research directly bearing on the subsequent development: Section 2.1 addresses methods for eliciting form preference only; Section 2.2 discusses eliciting both form preference and overall preference; and

Section 2.3 reviews optimization and machine learning algorithms for preference elicitation generally and conjoint analysis specifically.

2.1 Form Preference Modeling

Form preference has been addressed in engineering design using a variety of approaches, mainly differing in terms of the parametric nature of the preference function, and secondarily in terms of how these parameters are estimated (although we later discuss estimation extensively for the real-time adaptive portion of our implementation, we will be largely agnostic regarding estimation technologies otherwise). Table 1 summarizes previous research focused on eliciting form preference via parametric models, primarily from the engineering design field.

[Table 1, "Parametric Models for Eliciting Form Preference", about here[3]]

Among the first approaches to optimizing the visual design space was Lai, Chang, and Chang (2005), who tested 2D designs for a passenger car by having three professional product designers develop appropriate initial candidates, which were then broadened using images of 125 existing passenger cars, and subsequently culled by a panel of experts to 27 combinative designs for a (parametric) Taguchi experiment. Despite its pioneering nature in calibrating form preference, their study would be of limited interest to marketers, due to its lack of parametric heterogeneity, the non-adaptive nature of its querying strategy, and to a lesser extent its reliance on 2D models and a ratings-based conjoint approach. Lugo et al. (2012) designed a wheel rim, and Reid et al. (2013) a vehicle shape, using similar methods, with a linear preference function estimated via standard regression techniques, but also adaptively at the aggregate level for 2D designs. By contrast, quadratic preferences that allowed for internal extrema were applied by Orsborn et al. (2009) to 2D shape design, using a choice-based instrument at the individual level. Sylcott et al. (2013b) included interaction terms among attributes, and Kelly et al. (2011) allowed both quadratic preferences and potential interactions, although both were limited to aggregated inferences. Interaction terms are especially important for form optimization, since some consumer-valued qualities depend on multiple geometrical elements: for example, a product is only "compact" through an interaction among its dimensions.

There are several limitations in this line of engineering design research. First, as noted earlier, most has sidestepped preference heterogeneity, so that results would suggest a single "optimal" design suitable for the population as a whole. Second, with few exceptions this line of

[3] Numbered tables and figures not placed in-text appear at the end of the manuscript.

work has relied on non-adaptive (and sometimes non-choice-based) query designs, which are demonstrably outperformed by adaptive techniques (Toubia et al. 2003, Toubia et al. 2004, Abernethy et al. 2008), which we take up again in Section 2.3 and in our proposed method. Third, nearly all prior research in the area has relied on 2D product representations, although there are exceptions that did not explicitly calibrate a form preference model (e.g., Ren and Papalambros 2011, Reid et al. 2013). In the marketing literature, Kim, Park, Bradlow, and Ding (2014) emphasized the importance of different kinds of attributes and ran conjoint studies that involved product designs for both hedonic (e.g., sunglasses) and utilitarian (e.g., coffeemakers) products, using 2D designs from a candidate list of 20 possibilities, in a manner similar to Dotson et al. (2019). To control 3D rendering parametrically is a challenge for optimization algorithms (e.g., Hsiao and Liu 2002) and conjoint interface designers, as well as for participants who may find such representations cumbersome to navigate, despite their being perceptually essential.

More recently, machine learning architectures – for example, generative adversarial networks – have been employed to process of large corpora of 2D imagery, particularly for vehicle design (Pan et al. 2017, Burnap et al. 2019). This approach allows "learning" of essentially arbitrary degrees of complexity and interaction among the (2D) design elements, but depends on a supervised (i.e., labeled) set of stimuli to interrelate design elements and aesthetic preference. Such methods can scale to large stimuli sets; for example, Pan et al. (2017) modeled perceptions of four pairs of aesthetic design attributes (e.g., "Sporty" vs. "Conservative") for over 13,000 2D images of SUVs from 29 manufacturers, while Burnap et al. (2019) analyzed aesthetic image ratings from 178 consumers who participated in "theme clinics" to evaluate 203 unique SUVs. Labeled stimuli show great promise as a way to engage in pre-study image analysis. By contrast, our use of machine learning focuses on adaptively altering each respondent's 3D design in real-time, and to concurrently evaluate trade-offs between form and function.

2.2 Joint Modeling of Form and Overall Preference

Engineers, industrial designers, and marketers often face conflicting design choices due to trade-offs between form and function attributes. For example, consumers typically want products – cell phones, clothing, automobiles, etc. – that are trim (form) yet durable (function), or sophisticated (form) yet moderately priced (function), attributes with conflicting design imperatives. There is presently little formal research to guide this trade-off. Two papers other than the present one have specifically addressed it, both centering on vehicle design; critical differences in approach and data requirements among the three studies are summarized in Table 2.

[Table 2, "Approaches to Relating Form and Function Preferences", about here]

One approach is to simply measure form and overall preference separately, then knitting them together in a model-based manner. In this vein, Sylcott et al. (2013a) conducted three separate conjoint surveys: the first for form preference; the second for function preference (without price); and the last for overall preference using two meta-attributes (e.g., form and function, each with three levels like low, medium, and high). Despite its intricacy, this approach not only potentially exacerbates problems stemming from demographic differences among sampled groups, it precludes incorporating form into an *overall* preference model, leaving their relationship indeterminate. Specifically, zeroing in on the individual-level sweet spot in the joint space of product designs and traditional conjoint attributes is impracticable.

Dotson et al. (2019) model the effect of 2D images on product choice by accounting for the utility correlation between similar-looking images and augmenting the standard choice-based conjoint task with direct ratings of image appeal. "Form" is accommodated in overall preference by including (population) mean image appeal ratings a separate attribute in the product utility specification; and heterogeneity in image appeal is modeled via a multinomial probit where the utility error covariance depends on the correlations in consumer image ratings. Despite the inherent scalability of the approach, it relies on collecting additional ratings data, and assuming that the underlying Euclidean distance metric for control points on the 2D product (automobile) shape captures *consumers'* perceptions of design similarity. Our approach, by contrast, obviates the need for separate pictorial evaluations, that it's possible to capture "shape distance" using a single (distance) metric, or indeed presuming that 2D imagery suffices to represent product topology. Rather, "form" and "function" information is collected contemporaneously, from the same individuals, and thereby allows an assessment of design preference heterogeneity unrelated to preference scales, metrics, or pre-standardized imagery.

This hints at a deeper problem bedeviling design optimization methodology overall: reliance on large stimuli sets is exacerbated by the intrinsic nonlinearity of the visual design space. For example, a consumer who likes pick-ups and loves sports cars may dislike mid-sized sedans, despite their being "intermediate" in a canonical parameterization of "car shape space". That is, designers cannot determine a set of stimuli among which a target consumer has modest preference differentials and simply interpolate between them to determine a utility surface. Moreover, keeping target visual stimuli constant across respondents runs the risk of mismeasuring heterogeneity: ideally, each respondent should be able to veer into the region of the design space containing her

most preferred product configuration, as opposed to being imprisoned within the convex hull or along the simplex edge bordered by the pre-configured designs. Lastly, screen images used in prior research are necessarily two-dimensional and fixed, projections of the 3D designs that respondents must visualize and integrate to fully experience: Although this reduces both complexity in presentation and latency in administration, it substantially reduces realism and respondent involvement. Our methodology therefore works entirely with configurable, on-the-fly renderable, 3D product representations, with no prior or exogenous human evaluation.

2.3 Optimization and Machine Learning Algorithms

The approach developed here leverages advances in computing power, web-based query design, machine learning, and optimization to allow efficient, scalable, real-time form optimization, as opposed to simply choosing among a pre-determined set of (typically 2D) alternatives. For purposes of comparison, we summarize research according to two dimensions: estimation methods / shrinkage properties (Table 3), and adaptive query design (Table 4). Our coverage is again selective and deliberately concise. The interested reader is referred to Toubia, Evgeniou, and Hauser (2007; their Table 1 especially) for extensive background on methods; to Netzer et al. (2008) for general challenges in preference elicitation; to Toubia, Hauser, and Garcia (2007) for both simulation results and empirical comparisons among traditional and polyhedral methods; to Halme and Kallio (2011) for an overview of choice-based estimation; to Chapelle et al. (2004) for machine learning in conjoint; to Dzyabura and Hauser (2011) specifically in reference to adaptive questionnaire design; to Huang and Luo (2016) for "fuzzy" and SVM-based complex preference elicitation; and to Burnap et al. (2019) for supervised learning approaches to aesthetics.

[Table 3, "Estimation Methods for Preference Elicitation Models", about here]

Hierarchical Bayesian (HB) methods have long been popular for estimating partworths in conjoint, and serve to overcome sparse individual-level information by shrinkage towards a population-based density (Lenk et al. 1996, Rossi and Allenby 2003). Due to stability and its near-ubiquitous use in applications (e.g., Sawtooth), we adopt HB for estimating individual-level partworths of the overall preference model. Toubia et al. (2003) proposed a polyhedral method especially well-suited to metric paired-comparison query design; this has been extended to adaptive choice queries using classical and Bayesian approaches (Toubia et al. 2004; Toubia and Flores 2007). In a similar vein, Evgeniou et al. (2007) proposed a distinct (compared with HB) method for shrinking individual-level shrinkage, minimizing a convex loss function.

Cui and Curry (2005) and Evgeniou et al. (2005) extended the Support Vector Machine (SVM) – a popular machine learning algorithm used in classification problems – to conjoint applications. Specifically, Evgeniou et al. (2005) proposed an SVM mix that can accommodate parametric heterogeneity by shrinking individual-level partworths toward population-based values using a linear sum. Because it lowers computational cost dramatically compared to HB, at a comparable level of accuracy, we adapt this method for form preference as well as adaptive design for both form and overall queries. As detailed in Section 3, we couple this with a Gaussian kernel rank SVM mix to handle non-linear form preference.

[Table 4: "Adaptive Query Design Methods," about here]

Adaptive question design methods for conjoint analysis typically employ "utility balance" (Huber and Zwerina 2007, Abernethy et al. 2008), wherein profiles in each choice set have similar utilities based on partworths estimated from previous answers (Toubia et al. 2007a, p. 247); the approach is comparable to "uncertainty sampling" for query strategy in the machine learning field (Settles 2010). Previous research on adaptive querying, outlined in Table 4, has demonstrated that it generally outperforms non-adaptive designs, especially when response errors are low, heterogeneity is high, and the number of queries is limited (Toubia et al. 2007a).

Polyhedral methods (e.g., Toubia et al. 2003, 2004, 2007) typically select a next query by minimizing polyhedral volume longest axis length, thereby finding the most efficient constraints (i.e., queries) to reduce the uncertainty of feasible estimates (i.e., the polyhedron). By contrast, Abernethy et al. (2008) measure uncertainty as the inverse of the Hessian of the loss function and select the next query by maximizing its smallest positive eigenvalue. More recently, Huang and Luo (2016) – building on methods from Tong and Koller (2001), Lin and Wang (2002), and Dzyabura and Hauser (2011) – proposed an adaptive decompositional framework for high dimensional preference elicitation problems, using collaborative filtering to obtain initial partworths, must-haves and/or unacceptable features. They also leverage previous respondents' data, generating questions using fuzzy SVM active learning and utility balance, as we do.[4] Such an adaptive query design strategy, based on data from one or a few respondents' data, may not be efficient in the early

[4] Hauser and Toubia (2005) demonstrate that imposing utility balance can lead to biased partworths and ratios thereof, in a metric conjoint setting. While we cannot claim our forthcoming bi-level query task is immune to such biases, it consists of alternating (4-point, symmetric) ordinal and dichotomous choice questions, and we later empirically examine the impact of various cutoff values for the former, finding them to have minimal effect on predictive accuracy. It has been shown theoretically that active learning (adaptive sampling) requires a balance between exploitation and exploration (e.g., Osugi et al. 2005), while in Abernethy et al. (2008), utility balance exploits current model knowledge to choose samples with the most uncertain prediction.

steps of sampling. We therefore use an SVM mix for adaptive query design with modest computational cost for shrinkage, which samples more efficiently despite insufficiency of individual response data, as explained in detail in Section 3. Our approach is, of course, tailored to product form optimization and the trade-off between form and functional attributes, while Huang and Luo (2016) method was not designed to address product aesthetics or geometry.

3 Proposed Model

3.1 Overview

Here we propose and develop a new method aimed at adaptively measuring the "utility" associated with both design elements and traditional product (conjoint) attributes. At the heart of the method are iterative "bi-level" questions. A bi-level question consists of two sequential sub-questions, as shown in Figure 1a: one for form alone, the other for both form and function. Before delving into specifics of implementation, we must stress that they are exactly that: details that enable the method to work quickly and reliably with real subjects. The formal properties of the responses used here have been chosen to be amenable to scalability, for ease of respondent use, and due to the availability of well-vetted algorithms, and *not* because the method would not "work" with other scale types. That is, although the bi-level adaptive questioning method is novel, the overall set-up is 'modular' in the sense of being able to swap in other question types, as computational power allows, although their efficiency, scalability, and speed would need careful testing. Rather, the key goal, enabled by the bi-level query, is allowing respondents to seamlessly interact with a parametric embedding of the product's topology, to achieve real-time visual design optimization.

[Figure 1, "Iterative bi-level Queries and Design Changes", about here]

That said, for the form question, we utilize (as per Figure 1a) a standard anchored scale task. Specifically, we present two 3D vehicle renderings and ask "Which of the following styles do you prefer?" Responses are indicated on an ordered 4-point scale: "left one is much better," "left one is better," "right one is better," or "right one is much better." Four points were used to allow a moderate degree of preference expression over a binary choice task, but without exact indifference, which would provide little 'traction' for the forthcoming adaptive algorithm. Next, for the purchase question, we presented the previous 3D vehicle renderings again with "functional attributes," such as price and MPG. The respondent was then asked "Which car would you be more likely to buy?" and made a binary choice between the two presented vehicles. Such bi-level questions are repeated a specific maximum number of times, set by the analyst. The potential tendency for respondents to

maintain their choice on the form question for the purchase question, irrespective of the newly supplied functional attribute information, was controlled for by counterbalancing, that is, by switching the order of the two sub-questions from round to round, as indicated in Figure 1a.[5] After a choice was made by the respondent, the two designs were updated for the next round of (maximally-informative) comparisons, as shown in Figure 1b.

3.2 Scoring, Utility, and Updating Algorithm

Design or form preference of individual i stems from the *form model*:

$$s_i = S_i(\mathbf{x}) + \varepsilon_{si} \tag{1}$$

where s_i is the form score, S_i is a non-linear preference function, $\mathbf{x}$ is a vector of design features representing the form, and ε_{si} is an error process. Based on the form score, the overall preference for individual i is then given by the following (linear-in-parameters) utility model:

$$Y = U_i(s_i, \mathbf{a}) = \lambda_i s_i + \beta_i^T \mathbf{a} + \varepsilon_{yi}. \tag{2}$$

The vector $\mathbf{a}$ consists of binary dummy variables for function attribute levels (i.e., a three-entry binary vector for a four-leveled attribute); λ_i is the weight of the form score; β_i is the partworth vector for functional attribute levels; and ε_{yi} is associated error. In other words, Eq. (2) is a standard conjoint utility specification, including the form score ($\lambda_i s_i$), to be calibrated via Eq. (1) and weighted via λ_i, using Eq. (2). [Note that it is further possible to include interaction terms in the specification for $U_i(s_i, \mathbf{a})$, for example, between form (s_i) and various attributes within $\mathbf{a}$, although this can lead to many additional estimated parameters and identification difficulties. Because, as shown later, doing so did not improve accuracy in our application we do not raise this possibility explicitly again, referring generically to $U_i(s_i, \mathbf{a})$ as linear-in-parameters.]

The two preference models, Eq. (1) and (2), are updated iteratively in real time by bi-level questioning, as per Figure 1a. The process unfolds as follows:

- **Form question**: an individual makes a metric paired-comparison between two forms created by design variables, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. The preference model $S_i(\mathbf{x})$ in Eq. (1) is trained; then form scores, $s_i^{(1)}$ and $s_i^{(2)}$, are estimated. Finally, two function attributes (price and MPG in our application), $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$, are sampled for the subsequent purchase question.
- **Purchase question**: an individual makes a binary choice between two bundles of form and

[5] The actual interactive interface used for this study can be accessed at vehiclechoicemodel.appspot.com. Identifying information required by IRB has been removed for journal review.

functions $[s_i^{(1)}, \mathbf{a}^{(1)}]$ and $[s_i^{(2)}, \mathbf{a}^{(2)}]$. The weight of the form score λ_i and the partworths for functions β_i in Eq. 2 are estimated. Finally, two forms, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, were sampled for the subsequent form question.

The overall process for querying, sampling, and learning is summarized visually, in flow chart form, in Figure 2, and verbally as follows:

[Figure 2, "Overall process for querying, sampling, and learning", about here]

- **Start**. A new questionnaire is initialized when an individual accesses the website.
- **Step 1**: *Sampling form pair*. A pair of vehicle renderings is generated from the design space based on the current form preference model. The pair is such that their expected form *preference* is roughly equal, but their *shapes* differ maximally from one another and from all forms used before. If this is the first question for the current subject, a fixed form pair is used.
- **Step 2**: *Querying form question*. A metric paired-comparison response is received from the subject.
- **Step 3**: *Learning form preference*. A form preference model is trained based on previous form responses from this subject. Once the form model is learned (or updated if not the first round), the *form scores* of previously sampled vehicle renderings for the current subject are updated. Former subjects' form preference models are used for shrinkage.
- **Step 4**: *Learning overall preference*. Except for the first round, the overall preference model was adjusted based on the updated form scores. Former subjects' overall preference models were used for shrinkage.
- **Step 5**: *Sampling function pair*. Generate function attributes (e.g., price and MPG) for the current pair of vehicle forms, based on the updated overall preference model.
- **Step 6**: *Querying purchase question*. A (binary) choice is received from the subject once the function attributes are shown along with the forms.
- **Step 7**: *Learning overall preference*. Same as Step 4. [Steps 1 - 7 complete an odd-numbered round. Even-numbered rounds switch the order of the form and purchase questions, as elaborated in steps 8-12 below.]
- **Steps 8-12**: Same as Steps 1 (*Sampling form pair*); 5 (*Sampling function pair*); 6 (*Querying purchase question*); 2 (*Querying form question*); and 3 (*Learning form preference*). [If the iteration has reached its maximum round number, m, go to Step 13; otherwise, to 1.]
- **Step 13**: *Querying validation question*. Several validation sets are presented and used to check hit rate. [If all subjects have finished, go to Step 14; otherwise, wait.]
- **Step 14**: *Final learning form preference*. Finalize individual-level form preference models using all other subjects' results.
- **Step 15**: *Final learning overall preference*. Finalize individual-level overall preference models using all other subjects' results.
- **End**. Hit rate checked using responses to the validation questions.

Because the purchase question can be characterized as "forced binary choice", a natural question concerns whether this preferable to, for example, a three-option query that includes "no choice" or an "outside option". While this is possible, a pilot pretest suggested that some

participants (recruited through crowdsourcing) briskly dispensed with the survey by frequently selecting "no choice", although this may occur less with highly committed or remunerated participants. We also direct the reader to Brazell et al. (2006) for additional advantages, e.g., disentangling "which" option(s) consumers prefer from volumetric predictions regarding "whether" they will purchase at all. Appendix B provides amendments to Eq. (3) and (4) for the "no choice" set-up, which entail no changes in core implementation algorithms.

During each survey, we use the rank SVM mix algorithm (Evgeniou et al. 2005) for rapid training and an uncertainty sampling scheme similar to Settles (2010) and Tong and Koller (2002) for real-time generation of comparison pairs. We follow the suggestion of Toubia et al. (2013), "Once the data have been collected, we recommend combining our adaptive questionnaire design method with hierarchical Bayes, a well-proven method for estimating the parameters given the data". That is, after all user data are collected, we estimate individual partworths using standard hierarchical Bayesian (HB) techniques (Lenk et al. 1996, Rossi and Allenby 2003). We note that, unlike HB, SVM does not rely on an explicit notion of likelihood, although we show in Appendix C that it is consistent with hinge-loss and a Gaussian prior (similar to Evgeniou, Pontil, and Toubia 2007, p. 807). We emphasize again that the overarching method is agnostic on specific (machine learning) algorithms, which can be replaced with alternatives suited to the analyst's specific survey environment and research goals, dependent on computational speed and scalability (a topic we explore later empirically).

We next elaborate on how these algorithms are applied: Section 3.3 discusses learning methods for both the form and overall preference models, as well as the rationale for handing these separately, while Section 3.4 addresses sampling methods to generate pairs for comparison.

3.3 Learning Preferences

3.3.1 Form Preference Learning

As mentioned earlier, the form preference model in Eq. (1) is trained using a rank SVM mix algorithm. The idea is to fit a model consistent with the metric paired-comparison results, i.e., if one form of the pair is much preferred to the other, the preference gap between the two should also be larger than a pair that is less differentiated, i.e., one is preferred to the other. Following the treatments in Joachims (2002) and Chapelle and Keerthi (2010), the training problem can be formulated as follows:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T\phi(\mathbf{x}_j^{(1)}) - \mathbf{w}^T\phi(\mathbf{x}_j^{(2)}) \geq c_j, \forall j = 1, \dots, m \\ \text{where} \quad & c_j \in \{1,2\}. \end{aligned} \tag{3}$$

Here $\mathbf{x}_j^{(1)}$ and $\mathbf{x}_j^{(2)}$ are the design features for the chosen and unchosen forms in the j-th questions, respectively. User responses are represented by c_j: when $\mathbf{x}_j^{(1)}$ is "better" than $\mathbf{x}_j^{(2)}$, c_j is set to 1; when $\mathbf{x}_j^{(1)}$ is "much better", c_j is set to 2. The objective in Eq. (3) follows a hard-margin SVM formulation where $\mathbf{w}^T\mathbf{w}$ represents the model complexity. Later, in our simulation studies, we will explore both the sensitivity of the hard-margin SVM to choices of c_j, as well as its "brittleness" to contradictory preference ordering on the part of respondents. It is possible to view (3) in terms of its Lagrangian, $L(\mathbf{w}, \boldsymbol{\alpha})$; Appendix C details its relation to a negative log-likelihood with parameters $\boldsymbol{\alpha}$ balancing the Gaussian prior ($\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$) and the data (i.e., $\mathbf{x}_j^{(1)}$ preferred to $\mathbf{x}_j^{(2)}$).

One can project $\mathbf{x}$ to a high-dimensional space, i.e., $\mathbf{x} \to \phi(\mathbf{x})$, so that the constraints for $\forall j = 1, \dots, m$ can be satisfied. The dual problem of (3) can be expressed as follows:

$$\begin{aligned} \min_{\alpha} \quad & \tfrac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{c}^T\boldsymbol{\alpha} \\ \text{subject to} \quad & \boldsymbol{\alpha} \geq 0, \end{aligned} \tag{4}$$

where $\boldsymbol{\alpha}$ are Lagrangian multipliers and $\mathbf{Q}$ is an m by m matrix with each element, Q_{ij}, being the inner product $\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$. Following Chang and Lin (2011), a common choice for this inner product relies on the Gaussian kernel, which we denote as K:

$$K(\mathbf{x}_i, \mathbf{x}_j) \equiv \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle := \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2), \tag{5}$$

where the Gaussian parameter γ is set at $\gamma = 1/(\text{number of design features})$.[6] The dual problem can then be solved efficiently following the algorithm of Fan et al. (2005): based on the resultant Lagrangian multipliers, $\boldsymbol{\alpha}$, user preferences on a given form $\mathbf{x}$ can be quantified as

$$S(\mathbf{x}) = \sum_{j=1}^{m} \alpha_j (K(\mathbf{x}, \mathbf{x}_j^{(1)}) - K(\mathbf{x}, \mathbf{x}_j^{(2)})), \tag{6}$$

where $\mathbf{x}_j^{(1)}$ are all chosen forms during the survey and $\mathbf{x}_j^{(2)}$ the unchosen ones. For stability and comparability, the design features $\mathbf{x}$ are normalized to have zero mean and unit standard deviation

[6] Note that this does not uniquely specify ϕ, only the "inner product" function. In fact, defining ϕ is unnecessary here, since the dual problem to Eq. (4), owing to "strong duality", only requires the definition of the inner product $\phi^T\phi$ for the calculation of matrix Q.

before being used in training and prediction. Note that a soft-margin SVM formulation (Cortes and Vapnik 1995, Cristianini and Shawe-Taylor 2000) could be used in place of (4) to deal with 'noisy' user responses, a topic we examine later via simulation for the hard-margin SVM.

Due to the limited data collection from individuals, it is desirable to leverage data collected from *all* participants to improve the robustness and accuracy of individual-level preference models, similar to the shrinkage underlying HB methods. As discussed earlier, we use the (modest computational cost) method of Evgeniou et al. (2005), with partworth of participant i given as

$$\mathbf{w} = \sum_j \alpha_j \left(\phi(\mathbf{x}_j^{(1)}) - \phi(\mathbf{x}_j^{(2)})\right), \tag{7}$$

and population-wise partworth as

$$\bar{\boldsymbol{w}} = \frac{1}{N}\sum_{n=1}^{N} \sum_j \alpha_j^{(n)} \left(\phi(\mathbf{x}_j^{(1,n)}) - \phi(\mathbf{x}_j^{(2,n)})\right), \tag{8}$$

where N is the total number of participants, $\alpha_j^{(n)}$ is the Lagrangian multiplier for the j-th question for participant n, and $\mathbf{x}_j^{(1,n)}$ and $\mathbf{x}_j^{(2,n)}$ are the chosen and unchosen forms in that question, respectively. With a weighting factor $\eta_i \in [0,1]$, the individual form preference for participant i and a given form $\mathbf{x}$ can be calculated as:

$$\begin{aligned} S_i^*(\mathbf{x}) &= (\eta_i \mathbf{w} + (1-\eta_i)\bar{\boldsymbol{w}})^T \phi(\mathbf{x}) \\ &= \eta_i \sum_j \alpha_j \left(\langle \phi(\mathbf{x}_j^{(1)}), \phi(\mathbf{x})\rangle - \langle \phi(\mathbf{x}_j^{(2)}), \phi(\mathbf{x})\rangle\right) \\ &\quad + (1-\eta_i)\frac{1}{N}\sum_{n=1}^{N} \sum_j \alpha_j^{(n)} \left(\langle \phi(\mathbf{x}_j^{(1,n)}), \phi(\mathbf{x})\rangle - \langle \phi(\mathbf{x}_j^{(2,n)}), \phi(\mathbf{x})\rangle\right) \\ &= \eta_i S_i(\mathbf{x}) + (1-\eta_i)\frac{1}{N}\sum_{i=1}^{N} S_i(\mathbf{x}) \end{aligned} \tag{9}$$

We note that Eq. (9) is used only for the adaptive query design, and not for coefficient estimation (which is fully Bayesian), so values of η_i primarily affect generation efficiency. Specifically, if η_i is small, the function of individual i shrinks strongly toward the population-level function. For active learning during the survey, η_i can be selected at the discretion of the analyst: when there are few prior respondents, a large value of η_i can be used; otherwise, a smaller η_i value is appropriate. For estimation of form preference after finishing the survey, optimal η_i for the final estimation can be determined by cross-validation. That is, when we use J questions, $J-1$ responses are used for training and the remaining response is used for assessing prediction performance. J rounds of cross-validation are performed with different test data to minimize prediction error. In

our experiments, we used $\eta_1 = 1$ and $\eta_N = 0.7$ for the first respondent and the last respondent, respectively, with intermediate respondents interpolated linearly between these values.

3.3.2 Overall Preference Learning

The coefficients in the overall preference model of Eq. (2) can be estimated analogously to those of the form preference model, that is, using a rank SVM mix algorithm for active learning during surveys and HB for population-level modeling. The problem formulation for individual-level learning is as follows:

$$\begin{aligned} \min_{\mathbf{W}_i} \quad & \mathbf{W}_i^T \mathbf{W}_i \\ \text{subject to} \quad & \mathbf{W}_i^T \mathbf{X}_{ij}^{(1)} - \mathbf{W}_i^T \mathbf{X}_{ij}^{(2)} \geq 1, \forall j = 1, \dots, m, \end{aligned} \tag{10}$$

where $\mathbf{W}_i^T = [\lambda_i, \beta_i^T]$ are the linear coefficients, and $\mathbf{X}_{ij}^T = [s_{ij}, \mathbf{a}_{ij}^T]$ consists of the form score and the binary dummy variables of the function attributes for the j-th comparison for individual i. Just as for (4), we apply the Fan et al. (2005) algorithm to the dual of (10). All constraints are set to be greater than or equal to 1, as the comparison in this case is binary rather than metric. In Appendix D, we demonstrate that "1" can be changed arbitrarily to any positive value, and also write out the dual of (10) explicitly. This dual is solved using a linear mapping, i.e., $\phi(\mathbf{X}_{ij}) = \mathbf{X}_{ij}$, and the resultant individual-wise partworth vector $\mathbf{W}_i$ can be expressed as:

$$\mathbf{W}_i = \sum_{j=1}^{m} (\mathbf{X}_{ij}^{(1)} - \mathbf{X}_{ij}^{(2)})^T \alpha_{ij}. \tag{11}$$

In order to leverage population-level data, the linear shrinkage method of Evgeniou et al. (2005) is again used for individual-level partworths, $\mathbf{W}_i^*$; specifically,

$$\mathbf{W}_i^* = \eta_i \mathbf{W}_i + (1 - \eta_i) \frac{1}{N} \sum_{i=1}^{N} \mathbf{W}_i, \tag{12}$$

where η_i is the weight for individual i, and N is the number of individuals.[7] For population-level preference modeling, a hierarchical binary logit model (Rossi et al. 2005), with weakly-informative and zero-centered priors, is used for estimation. Specifically, at the upper level of the Bayesian model, we assume $\mathbf{W}_i$ to have a multivariate normal distribution,

[7] Shrinkage weights are constant across preference vector elements, so do not leverage possible covariances across them. Fixed SVM shrinkage is appropriate when finding the optimal variance-covariance matrix is not affordable for real-time query generation, and our operationalization is consistent with the standard LIBSVM package. Note that, for a linear function, partworth covariances can be optimized along with means; but for a nonlinear function that maps samples to an infinite-dimensional space, this is infeasible.

$$\mathbf{W}_i \sim N(\mathbf{0}, \Lambda). \tag{13}$$

At the lower level, choice probabilities take binary logit form:

$$\Pr(y_{ij} = 1) = (1 + \exp[\mathbf{W}_i^T(\mathbf{X}_{ij}^{(2)} - \mathbf{X}_{ij}^{(1)})])^{-1}, \tag{14}$$

where $\Pr(y_{ij} = 1)$ and $\Pr(y_{ij} = 0)$ denote the probabilities of selecting $\mathbf{X}_{ij}^{(1)}$ and $\mathbf{X}_{ij}^{(2)}$, respectively, for the j-th question of individual i.

3.3.3 Differences in Learning Algorithms and Rationale for Separation

It is important to note that the formulations are different for form preference and for overall utility: while the latter is governed by an identity feature function, for the former, a Gaussian kernel is applied to the geometric features of each form, where the geometric feature vector contains all pairwise distances from the control points that define the 3D geometries; these control points are further governed by the input variables in a nonlinear way. Due to the infinite-dimensional nature of the feature space induced by the Gaussian kernel, learning or modeling the covariance matrix of the partworths during querying is infeasible, and is an identity for shrinkage purposes. The "online" estimation of SVM model parameters is thereby based on both current and previous responses: rather than formulating and solving a large SVM problem with all responses considered, for computational feasibility, parameter estimates from each individual survey are weighted linearly, with earlier responses receiving lower weights. That is, a speedy heuristic approach is used during the survey and query generation, and a formal, computationally costly Bayes model for subsequent (offline) estimation.

3.4 Sampling Questions

We adaptively sample the next pair of forms or functional attributes based on two criteria. First, a profile pair comprising a question should have as near to the same utility as possible according to the current model. The second criterion is to maximize the minimum distance from existing data points, from both the current participant and all previous ones. The implementation is as follows: among the pair, the first form (or function attribute set) is sampled solely by the second criterion:

$$\begin{aligned} \max_{\mathbf{x}_1^{new} \in [0,1]^{19}} \quad & \min_j \left\| \mathbf{x}_1^{new} - \mathbf{x}_j^{old} \right\|^2 \\ \text{subject to} \quad & lb \le \mathbf{x}_1^{new} \le ub, \end{aligned} \tag{15}$$

where $\mathbf{x}_1^{new}$ is the first form (or function attribute set) alternative for the next question, $\mathbf{x}_j^{old}$ are the j-th form alternatives used in previous questions, and $m_{\mathbf{x}^{old}}$ is the number of form alternatives used previously.

Once $\mathbf{x}_1^{new}$ is sampled, its form preference value (or utility) can be calculated based on the current model. The second sample, $\mathbf{x}_2^{new}$, is obtained via optimizing a weighted sum of the two criteria:

$$\begin{aligned} \min_{\mathbf{x}_2^{new} \in [0,1]^{19}} \quad & v_1 \exp(-\|S(\mathbf{x}_1^{new}) - S(\mathbf{x}_2^{new})\|^2) + v_2(\|\mathbf{x}_1^{new} - \mathbf{x}_2^{new}\|^2 + \min_j \left\|\mathbf{x}_2^{new} - \mathbf{x}_j^{old}\right\|^2) \\ \text{subject to} \quad & lb \le \mathbf{x}_2^{new} \le ub, \end{aligned} \tag{16}$$

where $S(\mathbf{x})$ is the form preference model and v_1 and v_2 are the weights. Eq. (16) again balances two objectives: the preference values of the two new designs should be similar, and the second design should differ from the first. By construction, $\mathbf{x}_2^{new}$ should be far away from not only previous samples, but also from the current first sample; v_1 and v_2 (or, equivalently, their ratio) are chosen by the experimenter to balance the two criteria (in our experiments, $v_1 = 0.99$ and $v_2 = 0.01$; Appendix E provides details on setting these values and results of using others). Due to high potential nonconvexity, locating each successive form pair is accomplished via genetic algorithms (GAs), by enumerating all combinations of attribute levels using Eq. (15) and (16).[8]

In simple terms, the active learning algorithm generates designs that are different from one another (and from existing designs) while being are similar in preference (according to the current model). These two principles serve the purposes of exploitation and exploration, respectively, which are commonly enacted in active learning algorithms, e.g., Tong and Koller (2001), Osugi et al. 2005), and Abernethy et al. (2008). We note that it is possible to optimize both questions in the bi-level query pair simultaneously rather than individually in a sequence, as per Eq. (15) and (16). A pragmatic challenge in doing so is computational speed of the application engine (exogenously set to 30 seconds in our Google-based implementation), and user tolerance for waiting, which is

[8] Details for the GA implementation are: for Eq. (15), population size = 20, max. generations = 100; for Eq. (16), population size = 50, max. generations = 500. Each iteration generates a set of parent pairs using the current population, where parent set size = population size. Pair generation is via a tournament scheme: two tournaments are played, each with set of (population size - 1) players uniformly randomly chosen from the population, and the "most fit" chosen as the parent. This pair then goes through a one-point crossover procedure, with cutting point uniformly randomly chosen: if the two parents are $x_1 = [x_{11}, x_{12}]$ and $x_2 = [x_{21}, x_{22}]$, the two output designs from crossover are $[x_{11}, x_{22}]$ and $[x_{12}, x_{21}]$, with a 0.1 mutation probability. The mutation operation picks one element of the vector x at a random location i (uniformly), and changes its value from x_i to $x_i + \delta$, where δ is uniformly drawn from $[-0.05, 0.05]$; the mutated value is bounded in [0,1]. The algorithm terminates when the maximum number of generations is reached.

considerably shorter. Sequential optimization / generation of the question pair is thereby a pragmatic compromise, not a theoretical limitation. Appendix F provides full details for simultaneous ("all-in-one") pair generation, as well as benchmarks relative to the sequential approach that support the superiority of the decompositional approach.

4 Application: Geometric Set-Up and Model Simulation

We demonstrate the benefits of the proposed approach by applying it to the design of a highly multiattribute, visually complex durable: a passenger sedan. As discussed in our overview of the literature, vehicle design has been among the main application domains of form design optimization models in the engineering discipline, and was proposed as a canonical application of the presentation of pictorial information in conjoint by Green and Srinivasan (1990).

To apply any method for form optimization requires a way to explore the space of designs. Both Green and Srinivasan (1990) and Dotson et al. (2019), as well as the overwhelming bulk of real-world applications in the marketing discipline, rely on a candidate set of images, which as discussed previously make interpolation or extrapolation precarious, along with parsimony challenges for heterogeneity estimation. An alternative approach common in engineering and some prior marketing research (e.g., Michalek et al. 2005) is an explicit product topology model, that is, a geometric representation of the external form and internal workings of the product. Here, our goal is less onerous, as form optimization only requires parameterizing the external (3D) shape of the product in question, not ensuring that, for example, it is possible to engineer *internal* components to conform with cost and safety constraints.

For vehicle form representation, we therefore developed a 3D parametric vehicle shape model (Ren and Papalambros 2011 provide additional technical detail; see as well Orsborn et al. 2009). This parametric model generates 3D renderings using two-leveled structures, as follows. First, nineteen design variables $\mathbf{x}$ (ranging from 0 to 1; see Figure 3) were determined to be sufficient to realistically set the coordinates of *control points*, which in turn generate (Bezier) surfaces of the 3D model. Examples of the underlying parameterization include the distance from the front grille to the midpoint of the hood, the elevation and pitch of the center of the windshield's highest point, etc. The full space of potential sedan shapes would doubtless require additional parameters – for example, door and window shape are not explicitly optimized – but the 19 variables used here provide a very wide array of configurations, covering a broad swath of tested models currently in the North American sedan market, and thereby provide a reasonable trade-off

between fidelity and parametric complexity / dimensionality. Figure 3 illustrates locations of all control points, some coordinates of which are determined directly by the design variables, whereas others were either fixed or adjusted automatically to maintain surface smoothness. During training, the set of 19 design variable values translate to 26 control points that map to some 325 (= 26×25/2) design features,[9] each representing the distance between a pair of control points.

[Figure 3, "19 Design Variables and Their Control Points", about here]

There is obviously a cornucopia of functional attributes important to potential car buyers. Here, we focus on two of the most critical, price and gas mileage (MPG), in terms of their trade-offs to one another and to the various form attributes embedded in the design model. In order to avoid presumptions of linearity of response, especially to price, we included five discrete attribute levels for vehicle price and MPG, selected based on sales data of so-called "CD cars" (passenger sedans) in the US. Specifically, we chose discrete points based on the 10th, 25th, 50th, 75th, and 90th percentiles of both price and MPG, as shown in Table 5. Given the coding of both price and MPG into binary level indicators, the model as implemented technically encompasses 8 binary attributes, as opposed to two used linearly in the "overall" preference utility function.

Table 5: Function Attribute Levels

Level	Price (MSRP)	MPG (city/highway)	Percentile (market data)
1	$23K	23/27	10th
2	$25K	23/29	25th
3	$26K	24/30	50th
4	$29K	25/31	75th
5	$31K	26/32	90th

Because the bi-level adaptive technique is novel, it was important to test it in theory before doing so in practice. Consequently, we first present an extensive series of simulations designed not only to demonstrate parametric recovery, but the fit and hit rate performance of each component of the bi-level querying technique.

[9] The full parametric model and mapping of the 19 variables to design features is available from the authors.

4.1 Design for Simulation and Empirical Application

As proposed at the outset, we presume that the analyst wishes to model both form and overall preferences at the individual level, and can pose but a limited number of questions via a single-shot survey instrument. Previous research (e.g., Sylcott et al. 2013a, Dotson et al. 2019) carrying out similar analyses do not lend themselves to such one-shot form vs. function preference assessments, due to the need for time-intensive analysis between the separate form and function survey instruments. To explore the benefits of overcoming this limitation, we simulate three possible modeling options, also estimated in the forthcoming empirical application, as shown in Table 6: Model 1 is the "base," Model 2 is the "half version" of the proposed model, and Model 3 the "full version." The "half version" allows assessment of the bi-level question structure and adaptive design separately.[10]

Table 6: Simulation models and characteristics

Models	Querying	Learning	Sampling
Model 1 (Base: single-level)	Single level: 20 purchase questions	Form and overall: HB (linear)	Non-adaptive (DOE)
Model 2 (Half: bi-level)	Bi-level: 10 form questions & 10 purchase questions	Form: Rank SVM mix (nonlinear) Overall: HB (linear)	Non-adaptive (DOE)
Model 3 (Full: bi-level & adaptive)	Bi-level: 10 form questions & 10 purchase questions	Form: Rank SVM mix (nonlinear) Overall: HB (linear)	Adaptive

To accord with typical real-world implementations, all three models are assumed to be informed by the survey responses of 100 subjects, each with a total of 20 questions, including form and purchase questions; an online pilot study suggested that 20 questions sufficed for this particular application. For validation, we used 100 hold-out questions each for form and purchase questions (i.e., 200 total) to compute hit rates. As described earlier, "form" consisted of 19 continuous design variables, whereas "function" comprised the five levels each for price and MPG.

[10] Because the proposed active learning algorithm was built specifically for a bi-level structure and sequential processing, Table 6 lacks a "single level & adaptive" option, so does not have 2×2 factorial structure. To enable comparison, we customized the adaptive sampling method to the single-level case, and report these results ("Model M1a") in Table 8 as well. Because they rely on a custom algorithm, we do not discuss them further.

In line with conventional (non-adaptive) conjoint analysis techniques, Model 1 ("base") used purchase questions only, and accommodated both form and function within a single linear preference model, which was estimated using standard HB methods. For DOE (design of experiments), a Latin hypercube sampling method was used to generate questionnaire designs for both continuous and discrete variables.[11]

Model 2 is the "half version" of the proposed model, and allows testing of the bi-level structure. The bi-level structure makes it possible for form and overall preference to be modeled using different specifications – specifically, a nonlinear model for form preference (Gaussian SVM mix model) and a linear model for overall preference), after which the form model can be nested into the overall model.[12] Form preference was estimated by a rank SVM mix (Gaussian nonlinear), overall preference using HB (linear), and a Latin hypercube sampling method was used for DOE. Compared to Model 1, Model 2 sacrifices 10 purchase questions while adding 10 separate form questions.

Model 3 is the full version of the proposed model, and tests the bi-level structure, non-linear specification, and adaptive questionnaire design effects together. The querying and learning structures are the same as Model 2, but Model 3 uses adaptive sampling so that form and function profiles are generated in real time and that each question has different form profiles (i.e., a potentially limitless number of forms).

Broadly speaking, the simulation design adapted those used widely in academic research applying conjoint methods (e.g., Arora and Huber 2001, Toubia et al. 2004, Evgeniou et al. 2005). A mainstay of previous research is that response accuracy is controlled by the magnitude of an individual's parameters (partworths), while respondent heterogeneity is controlled by the variance of parameters (across respondents). We operationalized accuracy and respondent heterogeneity by setting each to two levels, "low" and "high." For example, the magnitudes of parameters were set to $\beta=0.5$ and $\beta=3$ for low and high response accuracy, respectively. On a logit scale, these represent deviations in log-odds of 0.5 and 3.0 from a baseline of zero (i.e., $\beta=0$); or, in terms of probability,

[11] Throughout, for Latin hypercube sampling, we used the lhsdesign Matlab library; for HB, the hierarchical binary logit model in the rhierBinLogit R package (Rossi et al. 2005); and for rank SVM, we implemented the rank SVM algorithm based on the LIBSVM package (Chang and Lin 2011).

[12] For the forthcoming online experiment (Section 5), subjects may be more easily able to trade-off between form and function, because they are shown a pair of vehicle forms first, followed by price and MPG with the same forms, an empirical issue not easily addressed by simulation alone.

according to $(1 + \exp(-\beta))^{-1}$, which translates into 0.62 and 0.95, respectively, on a probability baseline of 1/2. The parameter variances were set relative to the level of β, to $\sigma^2 = 0.5\beta$ and $\sigma^2 = 3\beta$ for low and high respondent heterogeneity, respectively. Based on these parameters, four normal distributions were defined: β was drawn from each distribution, and then four partworth levels for each function attribute, $(-\beta, -\beta/3, \beta/3, \beta)$, were generated, keeping constant differences set to $2\beta/3$.

For creating individual form preference functions, 19 continuous design variables generated a complex form preference model via main-effects ("independent") parameters γ and "interaction" parameters δ. The independent term of the k-th design variable, γ_k was drawn from four pre-defined distributions (analogous to the method used for the function attributes). Specifically, for k = 1, 2, ..., 19, four points $(-\gamma_k/3, \gamma_k, -\gamma_k, \gamma_k/3)$ were generated, then cubic spline interpolation (denoted $\mathbf{\Phi}(\gamma_k, x_k)$) was applied to create a *continuous* function with respect to the k-th design variable, x_k. We then drew 19×18/2 = 171 interaction terms, δ_{ij}, representing the relationship between the i-th and j-th design variables (for $i \neq j$). The form function, nonlinear but continuous, is therefore:

$$S(\mathbf{x}) = \sum_{k=1}^{19} \mathbf{\Phi}(\gamma_k, x_k) + \sum_{i=1}^{19} \sum_{j=1}^{i-1} \delta_{ij} x_i x_j \quad (17)$$

The distributions of δ_{ij} were balanced in the sense that the independent and interaction terms were set to a 2:1 ratio. [Specifically, following Evgeniou et al. (2005), we randomly generated 1000 independent terms and 1000 interaction terms, then compared the ratios of absolute values of independent and interaction terms, stopping when the standard deviation of the normal distribution for δ accorded with the 2:1 ratio.] Form score weight, λ, in Eq. (2) represents the *importance* of form preference, and was selected to make the ratio of absolute values of form score s and function attribute preference $\beta^T \mathbf{a}$ to be 1:2 for the "low" and 2:1 for the "high" form importance cases. To do so, we generated 10000 random product profiles and 10000 consumer preference models, examined the ratio of absolute values of form score and function preference, then selected the values that allowed for the 1:2 and 2:1 ratios.

Consequently, we created eight consumer preference scenarios, as defined in Table 7 (note that form score weights are small because form score *values* are relatively large). To check hit rate, we generated five sets of all eight scenarios, so that 40 total scenarios were used for the simulation.

[Table 7: "Consumer Preference Scenarios," about here]

4.2 Simulation Results

Table 8 shows the results of the various simulation scenarios, where hit rates were taken as the mean across the five sets. An asterisk (*) indicates the best, or not significantly different from best at $p < 0.05$, across the three models.

[Table 8: "Simulation Hit Rates," about here]

Except for one case (low form importance, low response accuracy, and high respondent heterogeneity), the "full" Model 3 outperformed Model 1 for both form and overall hit rates. For the form hit rate, every case suggests that Model 2 (bi-level structure and non-linear modeling) offers sizable improvements over Model 1 (base). Every case also shows Model 3 performing as well as or better than Model 2 (adaptive vs. non-adaptive questionnaire design), significantly outperforming Model 2 in 5 out of 8 cases. For overall hit rate, half the cases favor Model 2 over Model 1. Model 3 performed as well as or better than Model 2 in all but one case, significantly outperforming Model 2 in 3 out of 8 cases. These simulation results suggest that the proposed bi-level adaptive method (Model 3) can handily outperform the conventional one (Model 1), even with the sacrifice of 10 purchase questions. Notably from the perspective of the goals of the present study, form preference accuracy can be improved substantially (increasing to 65.7% from a base of 52.1%, or a 26% improvement on average), enabling marketers to pass along more reasonable target design values to industrial designers and engineers.

We conducted several post-analyses to evaluate robustness of results to: number of questions, number of attributes, form preference accuracy, preference ordering inconsistency, and analyst-tuned parameters, as follows.

[Figure 4, "Sensitivity to the Number of Questions", about here]

Sensitivity to Number of Questions. We examine the effects of total number of form and purchase questions, from 10 to 60, in increments of 10 on hit rate. Results appear in Figure 4, using the results of Model 1 and Model 3 with what is arguably the most difficult scenario in Table 8: high form importance, low response accuracy, and high heterogeneity. Except for the 10-questions case (i.e., 5 form and 5 purchase questions for Model 3 vs. 10 purchase questions for Model 1), Model 3 consistently outperformed Model 1 in overall preference accuracy. This owes to the fact that the *form* preference accuracy (for hit rate) of Model 3 was always significantly better than for Model 1, *even though half of the purchase questions are sacrificed*. More overall questions enabled better form preference accuracy for Model 3, whereas the performance of Model 1 did not improve substantially after 30 questions.

Sensitivity to Number of Functional Attributes. The second post-analysis examined going from 2 to 10 functional attributes, again focusing on the high form importance, low response accuracy, high heterogeneity scenario. Although including 10 attributes is sometimes feasible (though highly taxing for respondents) in a real-world online study, the proposed query generation procedure would entail a higher-dimensional discrete optimization problem, potentially hindering real-time performance. We focus on Models 1 and 3 for a comparison across bi-level structure: results suggest going from 2 to 10 attributes has a small effect on form preference hit rate for Model 1 (52.3 to 51.5), but none for Model 3 (65.0 to 65.0). However, overall hit rate degrades sharply: Model 1, 87.2 to 51.5; Model 3, 88.7 to 54.6. Although this is necessarily speculative, form preference accuracy in Model 3 appears unaffected by number of functional attributes, owing to its form preference being trained separately using bi-level structure.

Form Preference Accuracy. Ideally, one would also like to assess parametric recovery for the true vs. estimated form preference models. But this is not directly possible, because the true and estimated functions have different types and numbers of parameters: the former created by combining cubic curves and interaction terms, the latter an estimated SVM whose parameters are Lagrangian multipliers. However, it is possible to compare using hit rate as a "scale-free" measure of prediction performance. Specifically, we compare form importance between the true and estimated models across individuals by using RMSE, with form importance calculated as the ratio the *range* of form preference (value after multiplying by form score weight, $\lambda_i s_i$, in Eq. 1) to range of function preference. For brevity, we do this for the same "difficult" scenario in Table 8 as before (high form importance, low response accuracy, and high heterogeneity). RMSE values are 2.141 for Model 1, 0.149 for Model 2, and 0.119 for Model 3, with all pairwise differences significant at $p < .005$. As such Model 3 is very strongly favored over the other two, with very inferior form prediction performance for Model 2.

Robustness to Respondent "Noise" in Preference Ordering. Because the hard-margin SVM lacks a dedicated error structure, one might presume it could "break down" in the case of contradictory preference orderings, for example, a consumer who effectively asserts A > B and, elsewhere, B > A. To assess this possibility, we conducted additional simulations adding a 'noise factor', i.e., probability that the customer reverses preference orderings. Specifically, for each form question, contradictory preference ordering occurs randomly, with probability 10% or 20%, and we compared both form and overall preference hit rate with the base (0%, i.e., non-contradictory) case, as follows. Overall, there is a moderate fall-off in form preference hit rate, but very little in overall preference. For Model 3, form preference hit rate is {65.1, 62.5, 58.6} for {0%, 10%, 20%} contradictory preference order probabilities, while overall preference – which includes the effect of covariates – is affected far less, {88.7, 88.0, 87.8}. In short, the hard-margin SVM provided reasonable performance in the face of even moderate preference inconsistency.

Robustness to c_j. Lastly, as mentioned earlier, we examine robustness to the choice of "gaps" in Eq. (3), that is, c_j being set to 1 (better) and 2 (much better). In Appendix G, we explain how this sensitivity can be assessed using the Lagrange multipliers of the solution of the dual problem in Eq. (4), and carry out two simulations: increasing the values of c_j while holding them in fixed 1:2 ratio ({0.1,0.2} to {1000,2000}), and altering the ratio itself (1:2 to 1:10). Results suggest almost complete insensitivity of form preference hit rate across all these scenarios.

4.3 Bi-Level Query Performance

The bi-level query method offers two advantages. First, from a learning (parameter estimation) perspective, note that while form preference is a nonlinear function on a high-dimensional parameter space, overall utility can be expressed as a linear-in-parameters function of form preference and functional attributes (as well as interactions, although we tested for and did not find these for our application). By using SVM for form preference and HB for choice, the bi-level method exploits this structure, and should have better generalization performance than a single-level model (and this is consistent with our forthcoming application, e.g., Table 10, Model 1 vs. 2). Secondly, from a query perspective, the form-only queries are used to estimate form preference, which supports the choice of informative purchase query, and this salutary effect on query selection is reflected in the comparison between Models 2 and 3. And to reiterate, as shown in Appendix F, the performance of sequential query generation, vs. simultaneous, is superior for a fixed budget.

However, simulation on synthetic data and an empirical application, no matter how compelling, do not constitute a theoretical guarantee. To this end, in Appendix A, we present a detailed formal proof addressing when and why the bi-level query is superior. It relies on the computation of generalization error for the single- and bi-level cases, and has two overarching conclusions: (1) When the form signal-to-noise ratio $||\beta||_1/\sigma_s$ is sufficiently large, and the form-to-utility noise ratio σ_s^2/σ_y^2 is sufficiently small, the bi-level questionnaire is superior; (2) In the extreme case where form responses are noiseless ($\sigma_s^2 = 0$), the bi-level query is always better.

5 Online Experiment

The simulation spoke clearly to the advantages of bi-level adaptive querying. But, as the saying goes, what works well in theory may not do so in practice. To assess real-world performance of the bi-level adaptive technique for form preference optimization, we conducted three online surveys that correspond with the models simulated in Table 8. Three online groups were recruited through ClearVoice Research, a prominent online panel provider, and, to accord with the simulation scenarios, each comprised 100 subjects. Demographics were specified to match with the general US adult population of car-owning households; post-analysis confirmed the accuracy of recruitment.[13] A total of 20 questions were used for learning and 10 holdout questions (i.e., 5 form and 5 purchase

[13] Averages were as follows: 50.2 years age; 82% Caucasian; 81% suburban or small town; 95% high school; 69% some college; 58% working; 4% student; 15% homemaker; 23% retired; $58767 household income; 55% married; 4.3 family size; 2.6 children; 65% spouses employed. Full cross-classified categorical breakdowns are available from the authors.

holdout questions) used to check hit rates. Respondents completed the online task of their own volition, with no time limits, on devices of their choosing. The survey mechanism was implemented as follows. On the client side, JavaScript, WebGL and ThreeJS were used to enable real-time 3D model rendering and interaction through mainstream web browsers, with no additional software requirements. Critically, users were able to rotate the real-time-generated 3D images for each presented form before deciding on their responses. On the server side, Google App Engine was used for both executing real-time machine learning algorithms and for data storage. [See Appendix H for details on the query engine platform and resultant respondent response time tests.]

5.1 Parameters and Model Performance

All three models were estimated as described in Sections 2 and 3. Parameter estimates for form and function attributes appear in Table 9, and we discuss implications about their natural groupings after comparing relative performance quality.[14]

Table 9: Parameter Estimates and Summaries

	Form	**Price**					**MPG (city/highway)**				
	λ_i	$23K	$25K	$26K	$29K	$31K	23/27	23/29	24/30	25/31	26/32
Mean	4.45	1.06	0.42	0.21	-0.83	-0.86	-0.95	-0.50	0.13	0.64	0.68
StdErr	0.25	0.14	0.08	0.07	0.11	0.11	0.10	0.08	0.07	0.07	0.08
Heterog.	1.66	0.89	0.46	0.36	0.76	0.77	0.53	0.37	0.26	0.40	0.41
r with λ_i	---	-0.22	-0.39	-0.45	0.35	0.35	0.40	0.34	0.10	-0.46	-0.44

Table 9 lists parameter means, standard errors, estimated random coefficients standard deviation (i.e., of the heterogeneity distribution), and random effects correlation (*r*) with the Form score. The Form score is clearly "significant", on average, and examination of the posteriors for individual λ_i suggests form is significantly (.05) positive for more than ¾ of the participants. In other words, "Form Matters" not just as an overall parameter mean, but for most people individually. The last row suggests that – generally speaking – people – who value Form also "prefer" higher prices and lower MPG, both of which are consistent with a higher WTP overall, and perhaps a larger automotive budget. [Note that, because of the zero-mean scaling for the Price and

[14] As mentioned earlier, interactions can be included in the overall utility model. We performed an exhaustive search over the homogeneous model space – where our model is nested in those with interactions – and a targeted search, via Stan (https://mc-stan.org/), over the heterogeneous one. We saw no "significant" interactions, inputting Price and MPG either as binary levels (4 df each) or as values (1 df each). Specifically, for the latter, 95% HDRs are, with all variables mean-centered: Styling*Price: [-0.039, 0.502]; Styling*MPG: [-0.665, 0.370]; Price*MPG: [-0.028, 0.098]. Full interaction model details are available from the authors.

MPG partworths within-person, interpreting their significance levels and heterogeneity distributions is less straightforward than for Form.]

Hit rates for the three models in the online experiments are shown in Table 10, with the proposed model performing best across the board.

Table 10: Hit Rates in Online Experiment

	Form preference hit rate	Overall preference hit rate
Model 1 (Base: single-level)	54.40%	57.20%
Model 2 (Half: bi-level)	62.0%* (7.6%)	62.00% (4.8%)
p against M1	0.015	0.111
Model 3 (Full: bi-level & adaptive)	64.0%* (9.6%)	68.2%* (11.0%)
p against M1	0.003	0.001
p against M2	0.537	0.027

Figures in parentheses show percentage improvement over "base" Model 1

*Best, or not significantly different from best, at p<0.05, across all models

Model 2 (in Table 10) clarifies the effect of the bi-level structure, which entails substantial improvements in prediction, an increase of 7.6% and 4.8%, for form and overall preferences hit rates, respectively, compared to Model 1 (or 14.0% and 8.4% of their respective baselines). Model 3 (again in Table 10) shows the effect of the bi-level structure as before, but also of adaptive sampling. These results further suggest that Model 3 offers an increase of 9.6% and 11.0% for form and overall preferences hit rates, respectively, compared to Model 1 (or 17.6% and 19.2% of their respective baselines) and an increase of 2.0% and 6.2% compared to Model 2 (or 3.2% and 10.0% of baseline). The overall pattern of results suggests that adaptive sampling is useful to elicit both non-linear form preferences and linear overall preferences. Specifically, the bi-level structure appears to have affected predictive accuracy for form preference more than for overall preference; and adaptive sampling affected overall preference predictions more than those for form.

Although the overarching purpose of this study is to model both form and function preferences *together*, within the confines of a one-shot survey, and to measure the trade-offs among specific design variables and functional ones, we did test another model that did not incorporate form. Specifically, we removed form attributes from Model 1 to check overall preference prediction

based on functional attributes alone. In Model 1a, we trained the overall preference model using only the function attributes, price and MPG, then re-checked hit rate. The results were dramatic: the hit rate increases to 64.6%, from the 57.2% of Model 1 (or 12.9% of baseline). This suggests that *predicting overall preference by incorporating form design variables and function attributes within a single linear model may be suboptimal as a general approach*. Model 2 in fact shows slightly poorer performance in overall preference hit rate, as it sacrifices 10 purchase questions and instead models form preference. The proposed method, Model 3, by contrast, affords significantly better prediction (68.2%) for overall preference than Model 1a (64.6%).

5.2 Trade-offs between Specific Form and Function Attributes

An examination of parameter estimates and summaries in Table 9 suggests that Form is important, and nontrivially correlated with "function" attribute levels; but these are coarse metrics of their interrelation, and don't hint at what to actually produce for a heterogeneous market. In this vein, we first examine whether there appear natural groupings – that is, a segmentation – within the data in terms of the overall "weight" placed on form and on the two functional attributes (price and MPG). Given their within-respondent zero-sum scaling, Price and MPG importances can be calculated by the difference between the highest and lowest partworth; these values and $\{\lambda_i\}$ are averaged across MCMC draws to compute three deterministic values for each of the 100 subjects, who can then be clustered using (first) hierarchical and (subsequently) K-means methods according to their form (λ), price, and MPG importances. Typical metrics suggested four clusters fit the data best; both raw and standardized averages appear in Table 11.

Table 11: Clustering Based on Form, Price, MPG

	Overall	**Group 1**	**Group 2**	**Group 3**	**Group 4**
Size	**100%**	**24%**	**23%**	**14%**	**39%**
			Raw Differences		
Form	**4.45**	2.92	5.42	2.26	5.62
Price	**1.92**	4.09	2.69	0.71	0.55
MPG	**1.63**	2.24	2.00	2.00	0.92
			Standardized Differences		
Form	**0**	-0.92	0.58	-1.31	0.70
Price	**0**	1.33	0.47	-0.73	-0.83
MPG	**0**	0.67	0.40	0.40	-0.79

The four clusters can thereby be roughly interpreted as:

Group 1: **Price** and **MPG** are important (relative to Form)
Group 2: **All three** (Price, MPG, Form) are valued in balance
Group 3: **MPG** is important (relative to Price especially)
Group 4: **Form** is **very** important (relative to Price and MPG)

Of the four groups, the fourth is far more concerned with vehicle aesthetics than the other three, while the first group doesn't appear to value Form very strongly. In other words, *willingness to pay for vehicle form* is high in group 4 and low in group 1. Group 3, however, has relatively high WTP, since both Form and MPG are valued *relative* to Price.

[Figure 5, "Optimal Designs for Four Extracted Clusters", about here]

But we can, of course, do more: because the underlying model is literally built around the idea of real-time visual generation, it can render the car designs "most liked" by each of the four groups. These appear, from the sideways and head-on perspective, in Figure 5, i.e., the "forms" along with the "functions" most valued by each group. Although this amounts to visual inspection, note that the design for Group 3 resembles the Toyota Prius: it has relatively streamlined silhouette, e.g., the transition from the hood to the front windshield. This is consistent with Group 3's weighting MPG the most, especially in comparison with Price. Similarly, the design for Group 4 has lower profile than those of Groups 1 and 2, consistent with their finding form very important. While the designs themselves may not look radically different, this is to be expected, for several reasons. First, our parametric model does not incorporate every element of car design (e.g., door and window shape, metal vs. plastic exterior panels) nor common attributes (e.g., color, audio, interior configuration). Second, the extracted designs indeed span a rather large swath of the product topology space for *this particular type of car*, as can be gleaned from examining the range of design variables and control points (to which we return later). And finally, we must keep in mind that these four groups, while clustered according to preference, still do not have *identical within-group preferences*. That is, the rendered designs are for the "centroid" of each group; it is possible using the model to design for smaller groups, or even individuals, if the analyst so wishes.

Regardless, once the model is run and both form and function parameters are obtained, it is possible to extract much more than purportedly homogeneous segments that parcel consumers on their valuation of design *overall* (vs. price and MPG). Rather, the analyst can use individual-level price coefficients and weights on the underlying control points to infer which *specific design*

elements individuals, or groups, are willing to pay for. Product designers can then compare the cost of provision of those design elements – say, a more sloped profile that would provide less space for the engine compartment but could enhance aerodynamics – with WTP for that element, as well as any available demographics that could serve as a classic discriminant function or hierarchical model covariate. We conclude our discussion of the empirical conjoint data with an analysis of this sort: which design elements are associated with relatively high WTP, and do these vary substantially across the respondent pool?

5.3 Trade-offs between Specific Form and Function Attributes

To examine the relationship between functional attributes like price and *specific* form attributes is somewhat more complex than the usual trade-off computations that typically follow conjoint. This is because, while most functional attributes in conjoint are "vector" type – e.g., all else equal, it is better to get *higher* mileage and haggle for a *lower* price – this is seldom the case with design attributes. For example, one might like a large and highly angled windshield, but both size and pitch are self-limiting: no matter how much one might like more of them, each eventually veers into dysfunctionality. That is, design attributes tend to be of the U-shaped ideal-point type, with a respondent-specific internal "Goldilocks" maximum. Because we have calibrated an individual-preference model, it is not difficult to calculate, for each respondent, four quantities: a maximum (what degree of that element is liked best, contingent on all the other form elements being jointly optimized), its form "score" (as per Eq. 1 & 2), and the associated gradient and Hessian. The latter two can be quickly computed using numerical techniques, and for each respondent; and stable quantities were obtained for all four design variables.

Sensitivity to Design Variables. The diagonal components of the Hessian – which will all be negative for internal maxima – correspond to curvature or sensitivity: how (un)willing the respondent would be to give up one unit of that form attribute? [Recall that all form elements were normalized before optimization, rendering them comparable on a dimensionless scale.]

[Table 12: "Form vs. Function Trade-offs," about here]

For each of the original 19 design variables (see Figure 3), summary statistics for these sensitivities appear in Table 12, where it is apparent that there is a wide variance across both design attributes and people. *Generally speaking, "front and center" design variables were valued far more than those harder to see while driving*. For example, x8 – elevation of the central point where the back windshield meets the roof – appears to be the least sensitive design element: its median value was -0.169 (tabled value = -0.2), compared to the mean (across all design variables) of -17.2.

In simple terms, respondents were, on average, 100 times less sensitive to this design parameter than the others. However, sensitivity to the x8 design variable was quite heterogeneous, as evidenced by its *mean* value of -36.8;, suggesting strong skewness in the distribution of sensitivities (i.e., half of respondents are below 0.5% of the mean value). By contrast, six design attributes stand out in terms of high sensitivity to change, listed with their medians (see Figure 3): x1 (-33.2) and x2 (-36.9), the horizontal and vertical position of the midpoint the hood/windshield join; x9 (-36.3) and x10 (-38.3), the horizontal and vertical position of the midpoint of the hood front; x14 (-34.4), the lateral displacement of the hood/windshield join point directly in front of the driver; and x19 (-32.8), the outward displacement of the driver's side hood front.

Trade-offs Against Design Variables. However, examining sensitivities alone is merely suggestive: perhaps the consumers who are most sensitive to *specific* elements of design also have the lowest *overall* value for design, relative to price (or MPG). It is even possible that nearly all respondents, while responding to design changes *in and of themselves*, have no trade-off against "functional" attributes, especially price. As such, we wish to construct metrics for "Willingness To Trade-off" (WTT) for each design attribute vs. the two functional attributes (price and mileage) in our study, as well as the functional attributes against one another, a standard "WTP" calculation in traditional conjoint (Sonnier et al. 2007). This is made more complex by our having allowed for nonlinear response to the five levels of both price and MPG. So, for simplicity of presentation, we compute two values, akin to the interquartile range, for each consumer's utility function: the difference in the 25th and 75th percentile values. That is, the partworth differences for \$29K vs. \$25K and 23/29 MPG vs. 25/31; or, more simply \$4000 and 2MPG, which can then be standardized into willingness-to-trade-off \$1000 and 1MPG (by dividing by 4 and 2, respectively). Finally, the deterministic part of the utility model is given as $\lambda_i s_i + \beta_i^T \mathbf{a}$, so that we can unambiguously answer "what % change in each design variable within s_i maintains utility if price were either \$1000 higher (WTT Price, WTTP) or mileage 1 MPG better (WTT Mileage, WTTM)?" For each consumer, this entailed both the "form" model for s_i and the estimated value of λ_i, which measures overall design importance.

As a check, we first computed that mainstay of conjoint studies, the WTP for MPG, a standard trade-off between conjoint attributes. The literature reports a wide range of values, depending on which sort of cars are studied, the method of task (e.g., conjoint, purchase data, or field experiment), demographic composition of the respondent pool, and, critically, the range of prices and MPG values studied. Greene's (2010) review of this literature highlights findings from Gramlich (2010), who found that WTP in \$/mile (calculated based on an increase from 25 to 30 MPG and a gas price of \$2/gal.), was approximately \$800 for luxury cars, \$1480 for compact cars, and \$2300 for SUVs; but that this rose \$1430, \$2580, and \$4100, respectively, for a gas price of

$3.50/gal (in 2008 prices). In our study (last row of Table 12), for mid-priced sedans, median WTP in $/mile was $2410,[15] well within the latter reported range, lending external validity to our results.

Our goal was to quantify trade-offs between our 19 design variables and both price and MPG. To do so, we computed median (and mean) values for willingness to trade off a .01 deviation (from optimum) in each design variable against price (WTTP, in $1000) and mileage (WTTM, in 1 MPG); these were multiplied by 100 to reflect the relative size of the .01 deviation to the normalized unit scaling for each of the 19 design variables. Results appear in the last four columns of Table 12. Because the trade-off between price and MPG was $2410, we would expect the values for MPG to be about 40% of those for price (empirical values based on the means in the last row of Table 12 were calculated as 35.6% and 32.0% for medians and means, respectively). [We refer henceforth to median values, since the distribution across consumers is quite skewed for some design elements; .01 deviations from design optimum are given by 1/100th the values in the table.]

It is apparent that some design variables were far more valued than others, roughly tracking with the results for the Hessian (columns 2 and 3 of Table 12). For example, as before, x1 and x2 (horizontal and vertical position of the hood/windshield join midpoint) were each valued by approximately $50 for each .01 deviation from optimum (tabled values, 5.36 and 4.86, respectively, in $1000 units/100), and would correspond with mileage losses of .0152 and .0176 MPG each. These two (x1, x2) were not alone, as several of the design variables showed similar substantial sensitivity, with .01 changes corresponding to approximately $50 in price (e.g., x9, x10, x14, x19).

Using Table 12, one can tally up medians (i.e., summing columns 4 and 9, then dividing by 100) to calculate an "omnibus" value for design changes *overall*, based on .01 deviations from each consumer's optimum design along all 19 dimensions. Doing so translates into approximately $439 in WTTP and .157 miles in WTTM, both substantial values, given the small 1% deviations. Of course, the underlying choice model, based on form utility, is highly nonlinear, so one must take care in extrapolating such "local" results to the entire design space, where whole regions are likely to show little slope due to their being non-viable for particular consumers (e.g., designs they actively dislike). Furthermore, even with substantial heterogeneity, the availability of individual-level results means that – given the cost of production of different design elements and demographics for respondents – a manufacturer could roughly compute whether a certain *kind* of consumer would be willing to pay the added cost of a proposed design alteration, or whether it might be worth reduced fuel efficiency or trade-off against any traditional (functional) conjoint attribute.

[15] Specifically, in this case, Δ(partworths of $25K and $29K)/Δ(partworths of 25/31MPG and 23/29MPG), rendered in $1/mile. Sonnier, Ainslie & Otter (2007) provide additional detail on such calculations.

5.4 Form and Function Trade-Offs for Group-Optimized Designs

One advantage of the model, as detailed in Section 5.2 and illustrated in Figure 5, is the ability to form groupings based on form and function preference, and determine "high utility" – that is, visually appealing – designs for each group. But model output also allows specific form vs. function trade-offs to be assessed for each design; these are also broken out, by median, in Table 12. These not only help characterize each group's core trade-offs, but help determine whether there is substantial heterogeneity in terms of valuation of design element. Such an assessment is critical, for example, in the literature on component commonality in flexible manufacturing (e.g., Fixson 2007): groups with low interest in, or WTP for, a particular form or functional element can often be provided with a relatively high-quality version by re-using existing designs and benefiting from production economies-of-scale.

Table 12 lists WTT Price and Mileage for the various design elements by Group, as well as the trade-off between MGP and Price. Results indicate substantial heterogeneity, with Group 4 – for whom Form is very important overall – having highest WTP on average: roughly triple the entire-sample value (7.04 vs. 2.31), although they don't have highest WTP for every element. By contrast, Group 1 – who value Form little compared with Price and MPG – have low WTP for nearly all design elements, yet are quite sensitive, with WTPs of $600 and $870, respectively, across the range of design variables x9 and x10, which concern the forward and vertical displacement of hood front midpoint. While considering every such form trade-off for the four groups would take us far afield, the point is that designers can, for any given customer grouping, determine *which particular design trade-offs are "worth it" in terms of consumer group WTP*. Any set of candidate designs can be so assessed. Or, given a full-scale costing model, nonlinear optimization can determine the "maximally profitable" design, for a particular group, over the design space, just as in standard conjoint for traditional attributes. As an example of this last consideration, the model suggests that the four groups also have substantial heterogeneity in terms of WTP for additional gas mileage. As calculated earlier, the median value for the entire consumer set was $2410, but the last row of Table 12 suggests this is quite heterogeneous: $970, $1,730, $6,290, $3,110 for Groups 1-4, respectively. One might have guessed that the "cares about design" Group 4 had high WTP overall, but they are only about average in WTP for additional gas mileage. Rather, Group 1, as suggested by the initial clustering, is an outlier in this regard, with each mile per gallon valued at only $970, well under the actual savings over the typical lifetime of a car, and Group 4 at the other extreme, perhaps suggesting a distaste for combustion-based engines.

6 Conclusion and Future Directions

Preference elicitation is among the great success stories of experimental and statistical methodology addressing central problems in marketing. As evidenced by widespread adoption throughout the world over the last four decades (Sattler and Hartmann 2008), conjoint methods in particular can currently be deployed, using web-based tools, by practicing managers, with a low upfront burden in selecting optimal stimuli sets and backend estimation technologies. For example, Sawtooth's Discover allows product designers to specify attributes and levels, with subsequent "heavy lifting" – fashioning orthogonal designs, choice-based stimuli sets, and Bayesian estimation – handled seamlessly in the background. Yet even the best current implementations of conjoint founder on the shoals of visual design: while adjectival labels (e.g., sporty, bold, posh, etc.) and pre-generated 2D imagery can easily be included as categorical stimuli and covariates, and help directionally identify "what consumers are looking for," they neither allow consumers to converge on specific designs that appeal to them nor designers to focus solely on *the design space*, rather than pre-rendered descriptions or depictions of that space.

This paper proposes what we believe to be the first comprehensive approach to the visual design problem, one that leverages both the sort of product topology modeling common in engineering and rapid, scalable machine learning algorithms to interweave with state-of-the-art preference elicitation methods developed in marketing. The resulting hybrid, using bi-level adaptive techniques and manipulable, real-time rendered imagery, can be deployed using standard web-based protocols to zero in on each consumer's preferred product design along with the sorts of attributes used traditionally in conjoint. The approach eschews descriptions or pre-set depictions of any sort, allowing post-hoc processing of individual-level data to determine trade-offs between common attributes like product price and visual design elements, as well as against design overall.

Our empirical analysis focuses deliberately on automotive design, as this is among the most complex durables that a consumer ever purchases: it is high-involvement, requires many trade-offs, and choices are deliberative. Cars are, notoriously, among the most design-intensive of all products. To take but a few ill-fated examples, Ford lavished over $6Bn and several years on the design of its Mondeo "world car", Pontiac still stings over the Aztek, and the Edsel is the stuff of legend. Because the overwhelming majority of widely-deployed durable products involve computer-aided design (CAD), engineers can readily provide low-to-moderate dimensional product form representations for use in generating real-time 3D models for use within the method, with scant additional mediation by marketers or econometricians.

There are several ways to broaden the present approach, and perhaps to achieve even greater scalability. In the former category is including dedicated costing models and exogenous constraints on production feasibility, both of which are critical in real-world design. Michalek et al. (2011) achieved the latter for a small durable in a heterogeneous market, and is in fact not especially difficult to incorporate into a machine learning approach, using both soft (cost) and hard (feasibility) constraints. That is, a designer or engineer could designate the subspace of "buildable" configurations and their attendant costs, so that either generated candidate 3D designs remained within its boundaries, or finalized products emerged from constrained or penalized optimization.

Our empirical example included a 19-dimensional design space that spanned a large swath of in-production consumer sedans, but hardly encompassed the full spectrum of passenger vehicles, let alone every aesthetic component thereof. While larger design spaces are possible, the main impediments are computational speed, efficient solution / generation of designs, and consumer willingness to participate in long conjoint tasks. One possibility takes a cue from Toubia et al. (2013), e.g., a look-up table could allow for rapid, adaptive query generation, as opposed to dedicated query-generation algorithms operating over a continuous (form or function) space. While consumer fatigue is a real issue for large number of "optimized" attributes, it can be mitigated through various experimental design and fusion-based techniques (Molin & Timmermans 2009).

For all their efficiency, adaptive algorithms could lead – either as intermediate or final choices – to "undesirable" designs, e.g., too costly for the manufacturer or seemingly odd to the consumer, especially in very large design spaces. Such algorithms typically involve some degree of hand-tuning, but a promising approach uses machine learning tools to weed out designs far away from those that have been either produced, or liked, conditional on an appropriate training corpus (Burnap et al. 2019). Such an approach would also help avoid "wasting" questions on stimuli unlikely to be selected, over and above prior responses in the design survey being administered.

These are all implementable improvements that could lead to crowdsourced, real-time, manufacturer-feasible design optimization. The present method nevertheless demonstrates that judiciously-chosen query and estimation techniques, coupled with existing product topology models, render visual design feasible, without prior, analyst-supplied preconceptions of underlying design attributes.

7 References

Abernethy, Jacob, Theodoros Evgeniou, Olivier Toubia, J-P Vert. 2008. Eliciting consumer preferences using robust adaptive choice questionnaires. *Knowledge and Data Engineering, IEEE Transactions on* **20** (2) 145–155.

Arora, Neeraj, Joel Huber. 2001. Improving parameter estimates and model prediction by aggregate customization in choice experiments. *Journal of Consumer Research* **28** (2) 273–283.

Bloch, Peter H. 1995. Seeking the ideal form: product design and consumer response. *The Journal of Marketing* 16–29.

Brazell, Jeff D., Christopher G. Diener, Ekaterina Karniouchina, William L. Moore, Válerie Séverin, and Pierre-Francois Uldry. 2006. The no-choice option and dual response choice designs. *Marketing Letters* **17** (4) 255-268.

Burnap, Alex, John R. Hauser, and Artem Timoshenko. 2019. Design and Evaluation of Product Aesthetics: A Human-Machine Hybrid Approach. *Available at SSRN 3421771*.

Burnap, Alexander, Ye Liu, Yanxin Pan, Honglak Lee, Richard Gonzalez, and Panos Y. Papalambros. 2016. Estimating and exploring the product form design space using deep generative models. In *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pp. V02AT03A013.

Chang, Chih-Chung, Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2** (3) 27.

Chapelle, O, SS Keerthi. 2010. Efficient algorithms for ranking with SVMs. *Information Retrieval* **13** (3) 201–215.

Chapelle, Olivier, et al. 2004. A machine learning approach to conjoint analysis. *Advances in neural information processing systems*. 257–264.

ClearVoice. 2014. ClearVoice research. http://www.clearvoiceresearch.com.

Cortes, Corinna, Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* **20** (3) 273–297.

Creusen, Marielle EH, Jan PL Schoormans. 2005. The different roles of product appearance in consumer choice*. *Journal of product innovation management* **22** (1) 63–81.

Cristianini, Nello, John Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Cui, Dapeng, David Curry. 2005. Prediction in marketing using the support vector machine. *Marketing Science* **24** (4) 595–615.

Dotson, Jeffrey P, Mark A Beltramo, Eleanor McDonnell Feit, Randall C Smith. 2019. Modeling the Effect of Images on Product Choices. *Available at SSRN 2282570* .

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

Evgeniou, Theodoros, Constantinos Boussios, Giorgos Zacharia. 2005. Generalized robust conjoint estimation. *Marketing Science* **24** (3) 415–429.

Evgeniou, Theodoros, Massimiliano Pontil, Olivier Toubia. 2007. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science* **26** (6) 805–818.

Fan, R.E., P.H. Chen, C.J. Lin. 2005. Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research* **6** 1889–1918.

Feit, Eleanor McDonnell, Pengyuan Wang, Eric T Bradlow, Peter S Fader. 2013. Fusing aggregate and disaggregate data with an application to multiplatform media consumption. *Journal of Marketing Research* **50** (3) 348-364.

Feit, Eleanor McDonnell, Mark A Beltramo, Fred M Feinberg. 2010. Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Science* **56** (5) 785–800.

Fixson, Sebastian K. 2007. Modularity and commonality research: past developments and future opportunities. *Concurrent Engineering* **15** (2) 85-111.

Gramlich, Jacob, 2010. Gas prices, fuel efficiency, and endogenous product choice in the US automobile industry. Working paper, Georgetown University.

Green, Paul E, Venkat Srinivasan. 1990. Conjoint analysis in marketing: new developments with implications for research and practice. *The Journal of Marketing* **54** (4) 3-19.

Green, Paul E, Abba M Krieger, Yoram Wind. 2001. Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, **31** (3) S56-S73.

Greene, David L. 2010. How consumers value fuel economy: A literature review. No. EPA-420-R-10-008.

Halme, Merja, Markku Kallio. 2011. Estimation methods for choice-based conjoint analysis of consumer preferences. *European Journal of Operational Research*, **214** (1) 160-167.

Helfand, Gloria, Ann Wolverton. 2009. Evaluating the consumer response to fuel economy: A review of the literature. National Center for Environmental Economics Working Paper 09-04.

Hsiao, S.W., Liu, M.C., 2002. A morphing method for shape generation and image prediction in product design. *Design studies*, **23** (6) 533-556.

Huang, Dongling, and Lan Luo 2016 Consumer preference elicitation of complex products using fuzzy support vector machine active learning." *Marketing Science* **35** (30) 445-464.

Huber, Joel, and Klaus Zwerina. 1996. The importance of utility balance in efficient choice designs. *Journal of Marketing research* **33** (3) 307-317.

Joachims, Thorsten. 2002. Optimizing search engines using clickthrough data. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 133–142.

Kelly, Jarod C., Pierre Maheut, Jean-François Petiot, Panos Y. Papalambros. 2011. Incorporating user shape preference in engineering design optimisation. *Journal of Engineering Design* **22** (9) 627–650.

Kim, H.J., Park, Y.H., Bradlow, E.T. and Ding, M. 2014. PIE: a holistic preference concept and measurement model. *Journal of Marketing Research*, **51** (3) 335-351.

Kotler, Philip G, Alexander Rath. 1984. Design: A powerful but neglected strategic tool. Journal of Business Strategy **5** (2) 16–21.

Krishna, Aradhna, 2012. An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. *Journal of Consumer Psychology*, **22** (3) 332-351.

Lai, Hsin-Hsi, Yu-Ming Chang, Hua-Cheng Chang. 2005. A robust design approach for enhancing the feeling quality of a product: a car profile case study. *International Journal of Industrial Ergonomics* **35** (5) 445–460.

Landwehr, Jan R, Aparna A Labroo, Andreas Herrmann. 2011. Gut liking for the ordinary: Incorporating design fluency improves automobile sales forecasts. *Marketing Science* **30** (3) 416–429.

Lenk, Peter J, Wayne S DeSarbo, Paul E Green, Martin R Young. 1996. Hierarchical bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science* **15** (2) 173–191.

Lugo, José E, Stephen M Batill, Laura Carlson. 2012. Modeling product form preference using gestalt principles, semantic space, and kansei. *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, 529–539.

MacDonald, Erin, Alexis Lubensky, Bryon Sohns, Panos Y Papalambros. 2009. Product semantics and wine portfolio optimisation. *International Journal of Product Development* **7** (1) 73–98.

Michalek, Jeremy J, Fred M Feinberg, Panos Y Papalambros. 2005. Linking marketing and engineering product design decisions via analytical target cascading." *Journal of Product Innovation Management* **22** (1) 42-62.

Michalek, J. J., Ebbes, P., Adigüzel, F., Feinberg, F. M., & Papalambros, P. Y. 2011. Enhancing marketing with engineering: Optimal product line design for heterogeneous markets. *International Journal of Research in Marketing*, **28** (1), 1-12.

Molin, Eric JE, and Harry JP Timmermans. 2009. Hierarchical information integration experiments and integrated choice experiments. *Transport reviews*, **29** (5) 635-655.

Netzer, Oded, Olivier Toubia, Eric T Bradlow, Ely Dahan, Theodoros Evgeniou, Fred M Feinberg, Eleanor M Feit, Sam K Hui, Joseph Johnson, John C Liechty, et al. 2008. Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters* **19** (3-4) 337–354.

Gunay Orbay, Luoting Fu, and Levent Burak Kara. 2015. Deciphering the influence of product shape on consumer judgments through geometric abstraction. *J. Mechanical Design* **137** (8) 081103.

Orsborn, Seth, Jonathan Cagan, Peter Boatwright. 2009. Quantifying aesthetic form preference in a utility function. *Journal of Mechanical Design* **131** (6) 061001.

Orsborn, Seth, Jonathan Cagan. 2009. Multiagent shape grammar implementation: automatically generating form concepts according to a preference function. *Journal of Mechanical Design* **131** (12) 121007.

Osugi, Thomas, Deng Kim, and Stephen Scott. "Balancing exploration and exploitation: A new algorithm for active machine learning." In *Data Mining, Fifth IEEE International Conference on*, pp. 8-pp. IEEE, 2005.

Pan, Yanxin, Alexander Burnap, Jeffrey Hartley, Richard Gonzalez, and Panos Y. Papalambros. 2017. Deep design: Product aesthetics for heterogeneous markets. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1961-1970..

Qian, Yi, Hui Xie. 2013. Which brand purchasers are lost to counterfeiters? An application of new data fusion approaches. *Marketing Science* **33** (3) 437-448.

Reid, Tahira N, Bart D Frischknecht, Panos Y Papalambros. 2012. Perceptual attributes in product design: Fuel economy and silhouette-based perceived environmental friendliness tradeoffs in automotive vehicle design. *Journal of Mechanical Design* **134** (4) 041006.

Reid, Tahira N, Erin F MacDonald, Ping Du. 2013. Impact of product design representation on customer judgment. *Journal of Mechanical Design* **135** (9) 091008.

Ren, Yi, Panos Y Papalambros. 2011. A design preference elicitation query as an optimization process. *Journal of Mechanical Design* **133** (11) 111004.

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

Rossi, Peter E, Greg M Allenby, Robert E McCulloch. 2005. *Bayesian statistics and marketing*. J. Wiley & Sons.

Sattler, Henrik, Adriane Hartmann. 2008. Commercial use of conjoint analysis. In *Operations management in theorie und praxis*, pp. 103-119. Gabler.

Settles, Burr. 2010. Active learning literature survey. *University of Wisconsin, Madison* **52** 55–66.

Sonnier, Garrett, Andrew Ainslie, Thomas Otter. 2007. Heterogeneity distributions of willingness-to-pay in choice models. *Quantitative Marketing and Economics* **5** (3) 313-331.

Sylcott, Brian, Jonathan Cagan, Golnaz Tabibnia. 2013a. Understanding consumer tradeoffs between form and function through metaconjoint and cognitive neuroscience analyses. *Journal of Mechanical Design* **135** (10) 101002.

Sylcott, Brian, Jeremy J Michalek, Jonathan Cagan. 2013b. Towards understanding the role of interaction effects in visual conjoint analysis. *ASME 2013 International Design Engineering Technical Conferences*. American Society of Mechanical Engineers, V03AT03A012.

Sylcott, Brian, Seth Orsborn, and Jonathan Cagan. 2016. The effect of product representation in visual conjoint analysis. *Journal of Mechanical Design* **138** (10) 101104.

Tong, S., D. Koller. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research* **2** 45–66.

Toubia, Olivier, Theodoros Evgeniou, John Hauser. 2007. *Oxford Handbook of Innovation*, chap. Optimization-Based and Machine-Learning Methods for Conjoint Analysis: Estimation and Question Design. Springer, New York.

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

Toubia, Olivier, Laurent Florès. 2007. Adaptive idea screening using consumers. *Marketing Science* **26** (3) 342–360.

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

Toubia, Olivier, Eric Johnson, Theodoros Evgeniou, and Philippe Delquié. 2013. Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Science* **59** (3) 613-640.

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

Tseng, Ian, Jonathan Cagan, Kenneth Kotovsky. 2012. Concurrent optimization of computationally learned stylistic form and functional goals. *Journal of Mechanical Design* **134** (11) 111006.

8 Tables and Figures Not In Main Text

Table 1: Parametric Models for Eliciting Form Preference

Research	Parametric preference function	Parameter estimation	Hetero-geneity	Survey	Query design	Product	Repre-sentation
Lai et al. (2005)	S/N ratio	Taguchi	Aggregate	Rating	Non-adaptive	Vehicle	2D Silhouette
Orsborn et al. (2009)	Quadratic	BTL	Individual	Choice	Non-adaptive	Vehicle	2D Silhouette
Kelly et al. (2011)	Quadratic w/ interaction	PREFMAP	Aggregate	Rating	Non-adaptive	Water bottle	2D Silhouette
Lugo et al. (2012)	Linear	Regression	Aggregate	Rating	Non-adaptive	Wheel rim	2D Rendering
Reid et al. (2012)	Linear	Regression	Aggregate	Rating	Non-adaptive	Vehicle	2D Silhouette
Tseng et al. (2012)	Neural network	ANN	Aggregate	Rating	Non-adaptive	Vehicle	2D Silhouette
Reid et al. (2013)	Linear	BTL	Aggregate	Choice	Non-adaptive	Vehicle & carafe	3D Rendering
Sylcott et al. (2013)	Linear with interaction term	MNL	Aggregate	Choice	Non-adaptive	Vase & vehicle	2D Silhouette
Sylcott et al. (2016)	Linear	MNL	Aggregate	Choice	Non-adaptive	Knife	3D Printed
Pan et. al (2017)	Neural Network	Adversarial Training	Individual	Choice	Non-adaptive	Vehicle	Pixels (50K)
Burnap et. al (2019)	Neural Network	VB + Adversarial	Aggregate	Rating	Non-adaptive	Vehicle	Pixels (200K)
Dotson et al. (2019)	Linear	MNP	Individual	Choice + Rating	Non-adaptive	Vehicle	2D

Table 2: Approaches to Relating Form and Function Preferences

	Dotson et al. (2019)	Sylcott et al. (2013a)	This study
Survey	Two separate surveys (1) form: rating (2) overall: choice	Three separate surveys (1) form: choice (2) function: choice (3) overall: pairwise comparison	Bi-level questions in single survey (1) form: metric paired-comparison (2) overall: choice
Time delay between surveys	Yes	Yes	No (real time)
Query design	Non-adaptive	Non-adaptive	Adaptive
Preference function	Form: covariance structure Overall: linear	Form: quadratic Function: linear Overall: linear	Form: radial basis Overall: linear
Estimation	Form: Euclidian distance Overall: Bayesian	Bradley-Terry-Luce (BTL)	Form: Rank SVM mix Overall: HB
Heterogeneity	Individual	Individual	Individual
Product	Vehicle	Vehicle	Vehicle

Table 3: Estimation Methods for Preference Elicitation Models

Research	Method	Shrinkage
Lenk et al. (1996) Rossi and Allenby (2003)	Hierarchical Bayes	Yes
Toubia et al. (2003)	Metric paired-comparison analytic-center estimation	No
Toubia et al. (2004)	Adaptive choice-based analytic-center estimation	No
Cui and Curry (2005)	Support Vector Machine (SVM)	No
Evgeniou et al. (2005)	SVM mix	Yes
Evgeniou et al. (2007)	Heterogeneous partworth estimation with complexity control	Yes
Dzyabura and Hauser (2011)	Variational Bayes Active Learning with Belief Propagation	Yes
Burnap et al. (2016)	Restricted Boltzmann Machine and Convex Low-Rank Matrix Estimation	Yes
Huang and Luo (2016)	Fuzzy SVM estimation (Lin and Wang 2002 algorithm*)	Yes
This study	Form preference: Rank SVM mix Overall preference: Hierarchical Bayes	Yes

* Lin, Chun-Fu, and Sheng-De Wang. 2002. Fuzzy support vector machines. *IEEE transactions on neural networks* **13** (2) 464-471.

Table 4: Adaptive Query Design Methods

Research	Method	Sampling	Data used
Toubia et al. (2003)	Adaptive metric paired-comparison polyhedral question design	Minimize polyhedron volume and length of longest axis	Individual's prior responses
Toubia et al. (2004) Toubia, Hauser, and Garcia (2007)	Adaptive choice-based polyhedral question design	Minimize polyhedron volume and length of longest axis	Individual's prior responses
Abernethy et al. (2008)	Hessian-based adaptive choice-based conjoint analysis	Maximize smallest positive eigenvalue of loss function Hessian	Individual's prior responses
Dzyabura and Hauser (2011)	Active machine learning (Adaptive) + Variational Bayes	Maximize expected information gain (reduction in posterior entropy)	Synthetic + Individual Responses
Huang and Luo (2016)	Adaptive Fuzzy SVM; collaborative filtering	Tong and Koller (2001) + minimal ratio margin	Both individual's and others' prior responses
This study	Adaptive metric paired-comparison SVM mix question design	Minimize difference between utilities of new pairs and maximize Euclidean distance among all profiles	Both individual's and others' prior responses

Table 7: Consumer Preference Scenarios

Form importance	Response accuracy	Respondent hetero-geneity	Form score weight (λ)*	Form attribute coefficients: Independent terms (γ)	Form attribute coefficients: Interaction terms (δ)	Functional attribute partworths
Low	Low	Low	0.0043	N(0.5, 0.25)	N(0, 4.80)	N(0.5, 0.25)
Low	Low	High	0.0044	N(0.5, 1.5)	N(0, 13.7)	N(0.5, 1.5)
Low	High	Low	0.0028	N(3.0, 1.5)	N(0, 56.3)	N(3, 1.5)
Low	High	High	0.0057	N(3.0, 9.0)	N(0, 88.4)	N(3, 9.0)
High	Low	Low	0.0173	N(0.5, 0.25)	N(0, 4.80)	N(0.5, 0.25)
High	Low	High	0.0176	N(0.5, 1.5)	N(0, 13.7)	N(0.5, 1.5)
High	High	Low	0.0112	N(3.0, 1.5)	N(0, 56.3)	N(3.0, 1.5)
High	High	High	0.0230	N(3.0, 9.0)	N(0, 88.4)	N(3.0, 9.0)

***** Form score weights are small owing to scaling. "True form preference" was created by combining multiple cubic functions and interaction terms, while "true function preference" arises from a simple linear function. Because output values of form preference (form score) are far larger than those for function preference, a small form score weight was used for balance.

Table 8: Simulation Hit Rates

Simulation design			Form preference hit rate				Overall preference hit rate			
Form importance	Response accuracy	Respondent hetero-geneity	Model 1	Model 1a	Model 2	Model 3	Model 1	Model 1a	Model 2	Model 3
Low	Low	Low	50.8	53.4	65.2	**66.2***	90.5	90.6	91.9	**93.2***
Low	Low	High	51.0	51.1	**65.6***	**65.3***	91.6	91.7	91.7	90.1
Low	High	Low	52.0	53.2	63.3	**66.7***	92.7	92.6	**93.6***	**94.6***
Low	High	High	51.2	52.0	63.4	**65.2***	89.7	90.1	**92.3***	**92.8***
High	Low	Low	52.5	52.3	**65.1***	**66.1***	87.2	87.1	87.9	**90.1***
High	Low	High	52.3	52.4	**65.2***	**65.1***	87.2	87.6	**88.1***	**88.7***
High	High	Low	53.5	53.6	62.9	**66.3***	93.0	93.3	92.8	**94.4***
High	High	High	53.2	53.8	62.4	**64.7***	87.5	87.6	**88.5***	**89.8***

*Best, or not significantly different from best, at $p<0.05$, across all models

Table 12: Form vs. Function Trade-Offs

Design	Hessian at Max		WTT: Price ($1000; Median)					WTT: Mileage (1 MPG; Median)				
Vars.	Median	Mean	Overall	Group 1	Group 2	Group 3	Group 4	Overall	Group 1	Group 2	Group 3	Group 4
x1	-33.2	-48.3	**5.36**	0.67	2.40	4.87	11.71	**1.52**	0.64	1.36	0.69	3.76
x2	-36.9	-61.6	**4.86**	0.62	2.89	6.44	13.47	**1.76**	0.64	1.53	0.98	3.40
x3	-12.4	-27.5	**2.10**	0.20	1.24	2.93	5.49	**0.67**	0.20	0.48	0.40	1.70
x4	-6.1	-27.3	**0.86**	0.07	0.35	1.21	2.71	**0.31**	0.08	0.17	0.24	0.86
x5	-2.5	-13.4	**0.36**	0.04	0.21	0.55	1.41	**0.12**	0.04	0.10	0.08	0.39
x6	-3.6	-165.9	**0.67**	0.06	0.32	0.94	1.85	**0.20**	0.06	0.14	0.14	0.51
x7	-6.4	-9.3	**0.80**	0.12	0.42	0.58	2.64	**0.25**	0.10	0.23	0.18	0.81
x8	-0.2	-36.8	**0.02**	0.00	0.01	0.14	0.08	**0.01**	0.00	0.01	0.02	0.02
x9	-36.3	-52.5	**4.20**	0.60	2.65	3.55	13.46	**1.65**	0.58	1.59	0.63	4.12
x10	-38.3	-141.2	**5.30**	0.87	2.79	5.95	13.71	**1.81**	0.57	1.72	0.70	4.65
x11	-1.9	-3.4	**0.17**	0.03	0.16	0.44	1.16	**0.08**	0.02	0.08	0.05	0.36
x12	-11.3	-12.0	**1.41**	0.25	0.87	1.93	4.38	**0.56**	0.13	0.57	0.23	1.21
x13	-18.2	-27.8	**1.94**	0.38	1.25	1.91	7.21	**0.89**	0.36	0.74	0.30	2.02
x14	-34.4	-49.7	**4.92**	0.69	2.37	4.72	16.46	**1.76**	0.70	1.48	0.65	3.60
x15	-10.3	-26.2	**0.84**	0.27	0.76	1.23	5.14	**0.41**	0.19	0.39	0.12	1.10
x16	-17.7	-26.4	**2.25**	0.48	0.81	2.36	6.21	**0.78**	0.40	0.58	0.36	2.17
x17	-6.2	-19.4	**0.80**	0.21	0.55	0.73	3.34	**0.29**	0.15	0.26	0.09	0.85
x18	-17.8	-45.9	**2.64**	0.38	1.29	2.19	8.27	**1.12**	0.40	0.76	0.31	2.08
x19	-32.8	-53.1	**4.45**	0.70	2.36	4.76	14.95	**1.46**	0.53	1.18	0.64	4.36
Mean	**-17.2**	**-44.6**	**2.31**	**0.35**	**1.25**	**2.50**	**7.04**	**0.82**	**0.31**	**0.70**	**0.36**	**2.00**
MPG	---	---	**2.41**	**0.97**	**1.73**	**6.29**	**3.11**	1	1	1	1	1

Reported values for Willing To Trade-off Price (WTTP) and Willing To Trade-off Mileage (WTTM) refer to units of $1000 or miles, respectively, for a .01 change in the design variable, multiplied by 100. These thus provide a "local" approximation of the full range of the design variables, each of which was normalized to lie on a unit scale. Design variable descriptions appear in Figure 3. Specifically, the final line's median WTP in dollars for each additional MPG is $2410 (overall), and $970, $1,730, $6,290, $3,110 for Groups 1-4, respectively.

Figure 1: Iterative bi-level Queries and Design Changes

Figure 1a: Iterative bi-level questions

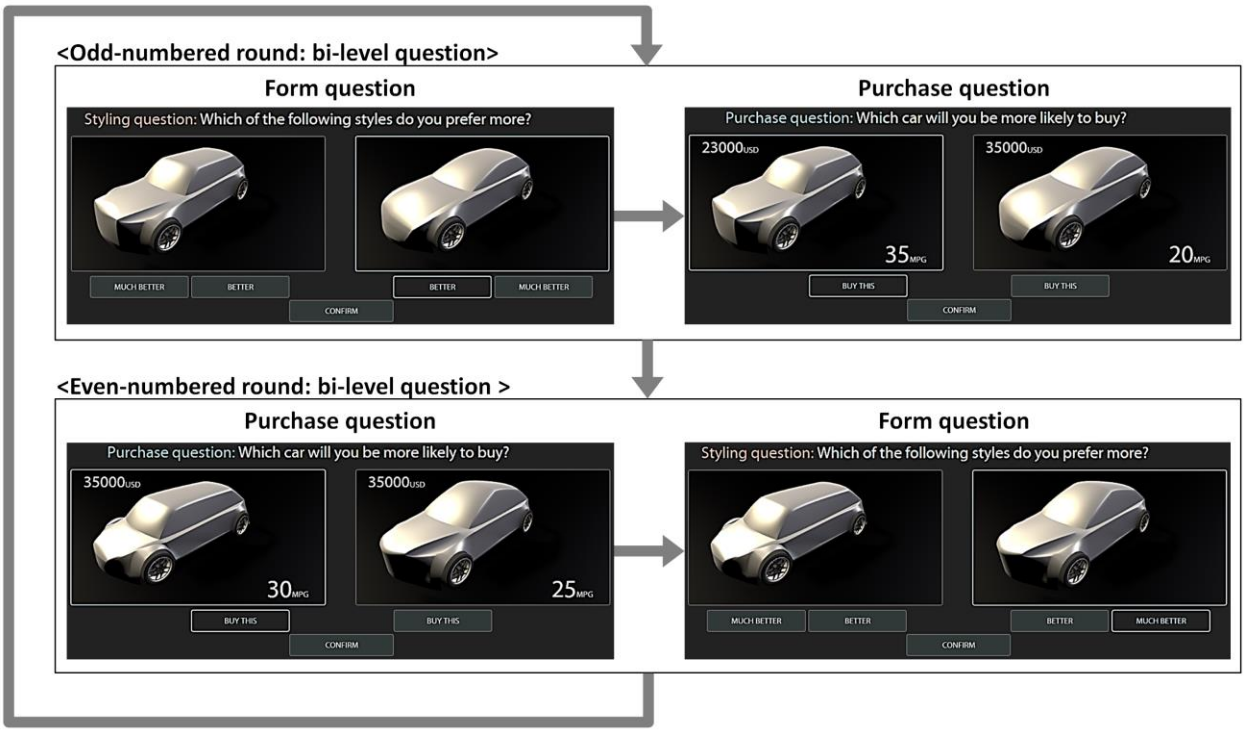


Figure 1b: Design Query Update Resulting from Prior Round Choice

Round 1:

Round 2:

Note: The "Round 2" designs result from answers to the "Round 1" bi-level queries; that is, both sub-questions of Round 1 lead to the design pair used in Round 2, which in turn lead to those generated for Round 3, etc.

Figure 2: Overall process for querying, sampling, and learning

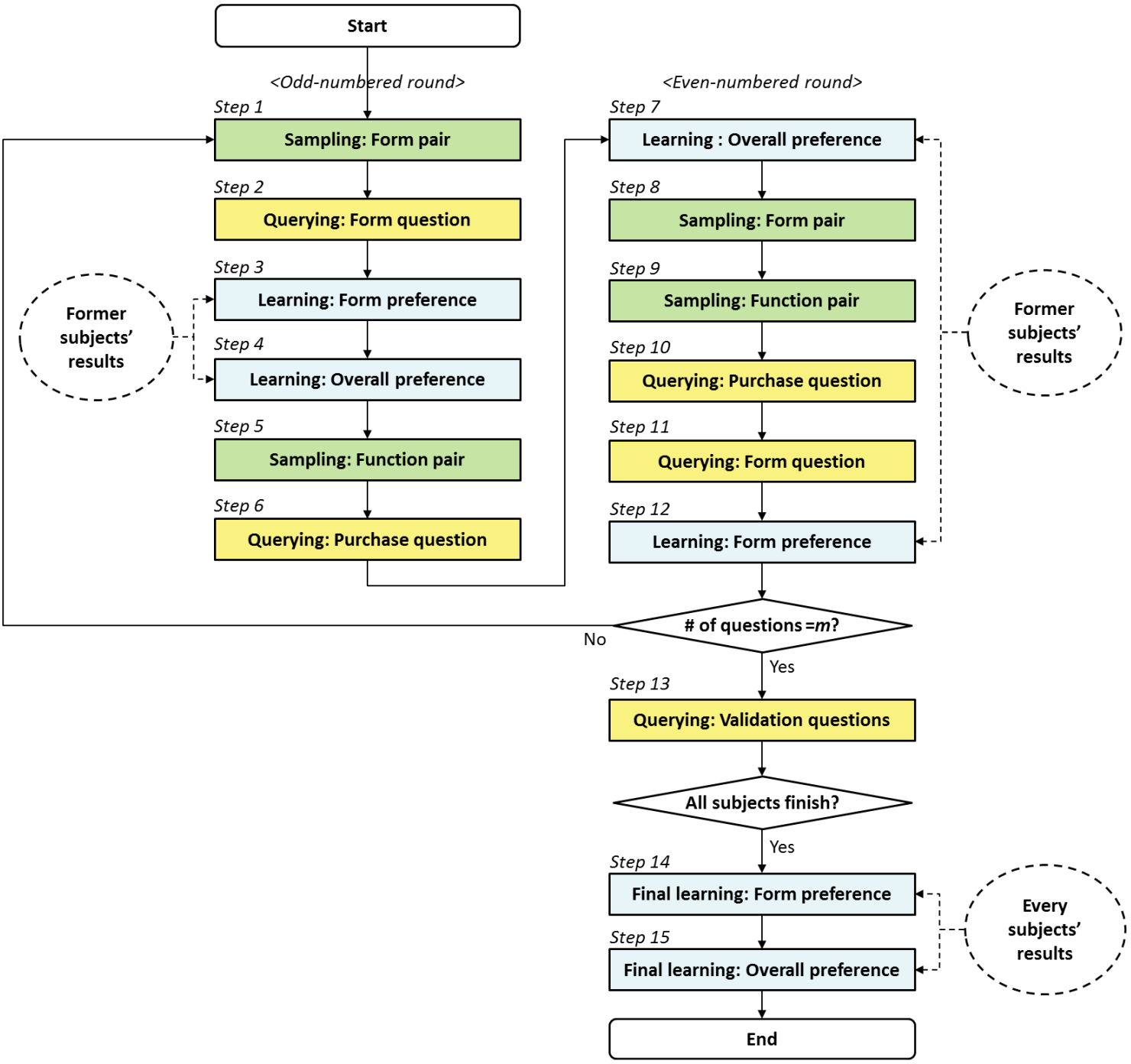
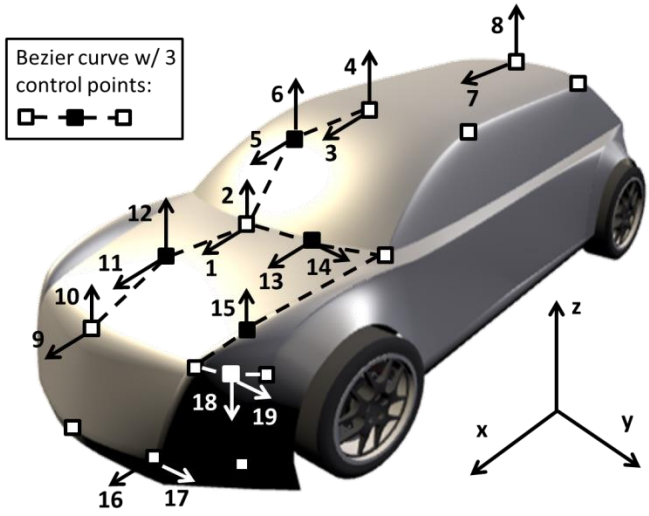


Figure 3: Nineteen design variables and Their Control Points



Legend:

x1, x2: forward (X) and vertical (Z) displacement of windshield / hood mid-join point
x3, x4: forward (X) and vertical (Z) displacement of windshield / roof mid-join point
x5, x6: forward (X) and vertical (Z) displacement of windshield curvature determination point
x7, x8: forward (X) and vertical (Z) displacement of hood / rear-window mid-join point
x9, x10: forward (X) and vertical (Z) displacement of hood front midpoint
x11, x12: forward (X) and vertical (Z) displacement of hood centroid point
x13, x14: forward (X) and lateral (Y) displacement of windshield / hood driver centerpoint
x15: vertical (Z) displacement of external hood edge curvature point
x16, x17: forward (X) and lateral (Y) displacement of front bumper lowest edge point
x18, x19: forward (X) and lateral (Y) displacement of hood / front bumper join point

Figure 4: Hit Rate Sensitivity to Total Number of Questions

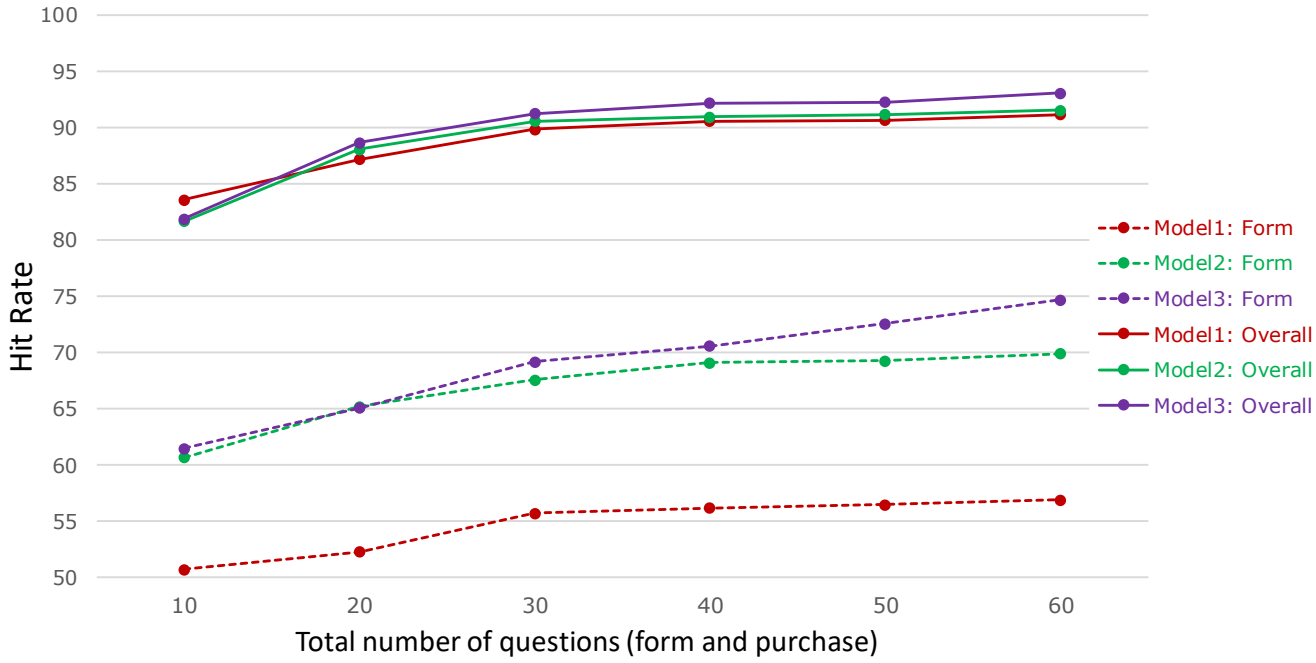
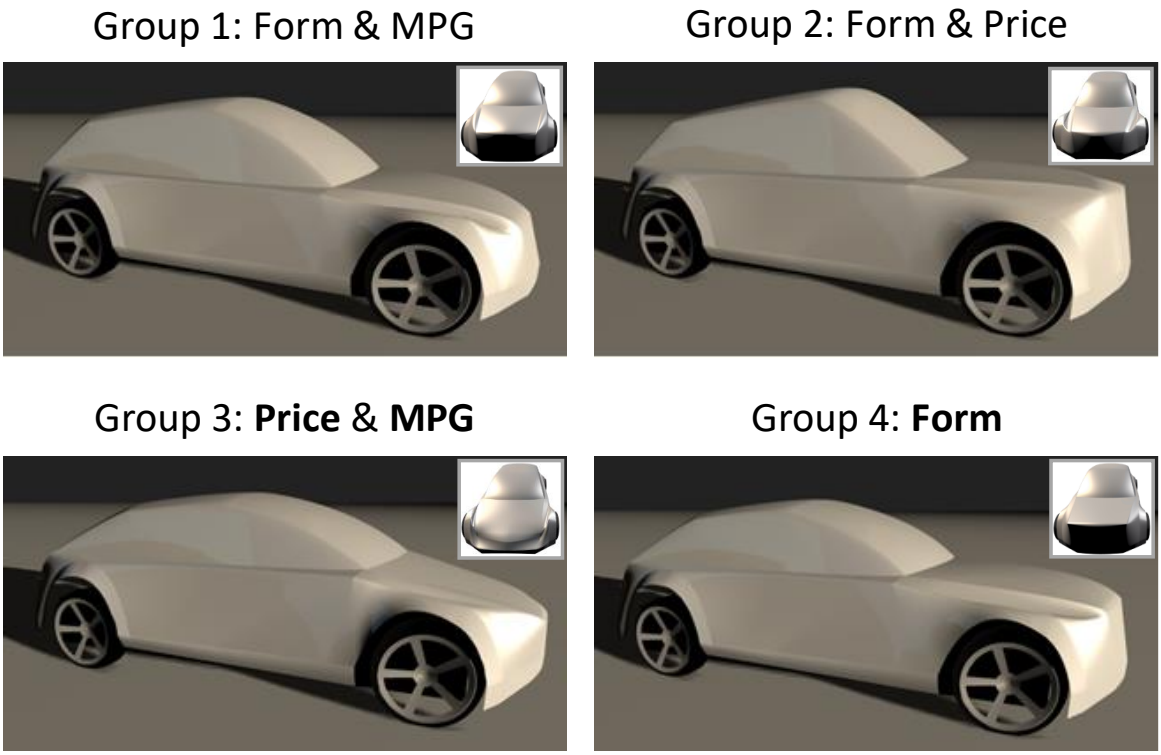


Figure 5: Optimal Designs for Four Extracted Clusters



APPENDICES: Technical and Online

MKSC-16-0281.R1

Form + Function: Optimizing Aesthetic Product Design via Adaptive, Geometrized Preference Elicitation

APPENDIX A: Superiority of Bi-Level Query Method

Online Only:

APPENDIX B: Accommodating the "No Choice" Option

APPENDIX C: Lagrangian Formulation of Eq. (3)

APPENDIX D: Insensitivity of Individual-Level Learning to Constraint Value

APPENDIX E: Setting Weights (v_1 and v_2) in Eq. 16

APPENDIX F: Simultaneous vs. Sequential Optimization for Eq. (15) and (16)

APPENDIX G: Robustness Checks for Preference Gap Cutoffs, c_j, in Eq. (3)

APPENDIX H: Details on Query Engine and Response Times

APPENDIX A: Superiority of Bi-Level Query Method

A1 Introduction

To provide insights on why the bi-level questionnaire works better, we introduce simplified data models for the single- and bi-level cases so that generalization error can be computed.

Specifically, we consider two linear models: The *form* model maps form features $x \in \mathbb{R}^{d_x}$ to the form score $s \in \mathbb{R}$, with parameters β and i.i.d. random error $\varepsilon_s \sim N(0, \sigma_s^2)$:

$$s = \beta^T x + \varepsilon_s. \tag{A1}$$

The *preference* model maps the form score s and product attributes $u \in \mathbb{R}^{d_u}$ to the utility $y \in \mathbb{R}$, with parameters b_s for the form score and b_u for the product attributes and i.i.d. random error $\varepsilon_y \sim N(0, \sigma_y^2)$:

$$y = b_s s + b_u^T x + \varepsilon_y. \tag{A2}$$

In the following, we discuss two cases corresponding to the two types of questionnaires studied in this paper: In the single-level questionnaire, we directly collect answers from purchase questions; and in the bi-level questionnaire, we collect answers for both form and purchase questions. For both cases, our goal is to build a predictive model so that for new inputs in the form of $\langle x, u \rangle$, we can predict their utility. We investigate the generalization errors in utility prediction using these two types of experiments.

As an overview, our insights are as follows:

- **When the form signal-to-noise ratio $||\boldsymbol{\beta}||_1/\sigma_s$ is sufficiently large, and the form-to-utility noise ratio σ_s^2/σ_y^2 is sufficiently small, the bi-level questionnaire is a superior choice.**
- **In the extreme case where form responses are noiseless ($\sigma_s^2 = 0$), the bi-level questionnaire is always better.**

A2 Case 1: Single-phase Questionnaire

A2.1 Data

In the first case, we consider the existence of data $\mathcal{D} = \{y^i, x^i, u^i\}_{i=1}^{N_1}$, i.e., we can observe the actual utilities, but not form scores. This is a simplification from the actual experiments, where only comparisons between ys rather than their actual values are observable.

We also make the following assumption to smooth the analysis: Let $X \in \mathbb{R}^{N_1 \times d_x}$ and $U \in \mathbb{R}^{N_1 \times d_u}$ be the data matrices where each row of X (U) is a data point $(x^i)^T$ ($(u^i)^T$). Denote $x_{\cdot j}$ ($u_{\cdot j}$) as the jth column of X (U). We assume that

- (**R1**) the matrix $D = [X, U]$ is column-wise orthogonal;
- (**R2**) $1^T x_{\cdot i} = 0$ for all i;

- (**R3**) $x_{\cdot i}^T x_{\cdot i} = N_1$ for all $i \neq j$.

(R1) can be achieved through design-of-experiments when $N_1 \geq d_x + d_u$, and is plausible when N_1 is large and x and u are independently sampled in each dimension. (R2) and (R3) can always be achieved by preprocessing the data to have zero means and unit variance column-wise. To further facilitate some of the detailed proofs, we also assume that each dimension of x and u is bounded (through truncated distributions).

A2.2 Prediction and uncertainty

For a specific dataset $\mathcal{D}$, we can estimate parameters for the following model through ordinary least square (OLS):

$$y = b_s \beta^T x + b_u^T u + \varepsilon_y + b_s \varepsilon_s. \tag{A3}$$

The model can be further simplified as

$$y = b^T w + \varepsilon, \tag{A4}$$

with $b^T := [b_s \beta^T, b_u^T]$, $w^T = [x^T, u^T]$, and the i.i.d. random variable $\varepsilon \sim N(0, \sigma_y^2 + b_s^2 \sigma_s^2)$.

Denote $\hat{b}_{\mathcal{D}}$ as the OLS estimate of b derived from data $\mathcal{D}$. We have $\mathbb{E}[\hat{b}_{\mathcal{D}}] = b$, and the variance-covariance matrix of $\hat{b}_{\mathcal{D}}$ (due to the noise in observations of y) is:

$$Var(\hat{b}_{\mathcal{D}}) = (\sigma_y^2 + b_s^2 \sigma_s^2)(D^T D)^{-1}. \tag{A5}$$

A2.3 Generalization error

The generalization error of $\hat{b}_{\mathcal{D}}$ is defined as

$$\begin{aligned} L(\hat{b}_{\mathcal{D}}) &= \mathbb{E}_w[(\hat{b}_{\mathcal{D}}^T w - b^T w)^2] \\ &= \int_w ((\hat{b}_{\mathcal{D}} - b)^T w)^2 p(w) dw \\ &= \int_w \left(\sum_{i=1}^{d_x+d_u} \Delta\hat{b}_{\mathcal{D},i} w_i\right)^2 p(w) dw \\ &= \int_w \left(\sum_{i=1}^{d_x+d_u} (\Delta\hat{b}_{\mathcal{D},i} w_i)^2 + \sum_{i=1}^{d_x+d_u} \sum_{j=i+1}^{d_x+d_u} (\Delta\hat{b}_{\mathcal{D},i} \Delta\hat{b}_{\mathcal{D},j} w_i w_j)\right) p(w) dw \\ &= \int_w \sum_{i=1}^{d_x+d_u} (\Delta\hat{b}_{\mathcal{D},i} w_i)^2 p(w) dw + \int_w \sum_{i=1}^{d_x+d_u} \sum_{j=i+1}^{d_x+d_u} (\Delta\hat{b}_{\mathcal{D},i} \Delta\hat{b}_{\mathcal{D},j} w_i w_j) p(w) dw \\ &= \sum_{i=1}^{d_x+d_u} \int_{w_i} (\Delta\hat{b}_{\mathcal{D},i} w_i)^2 p(w_i) dw_i + \sum_{i=1}^{d_x+d_u} \sum_{j=i+1}^{d_x+d_u} \int_{w_i,w_j} (\Delta\hat{b}_{\mathcal{D},i} \Delta\hat{b}_{\mathcal{D},j} w_i w_j) p(w_i, w_j) dw_i dw_j \end{aligned} \tag{A6}$$

where $\Delta\hat{b}_{\mathcal{D},i}$ is the ith element of $\hat{b}_{\mathcal{D}} - b$. We assume that each element of the test data point w_i is drawn i.i.d. with $\mathbb{E}[w_i] = 0$ and $Var(w_i) = 1$ (again, this can be done through proprocessing of the data space). Therefore, we have

$$\int_{w_i,w_j} w_i w_j p(w_i, w_j) dw_i dw_j = 0, \tag{A7}$$

and

$$L(\hat{b}_{\mathcal{D}}) = \sum_{i=1}^{d_x+d_u} \Delta\hat{b}_{\mathcal{D},i}^2. \tag{A8}$$

Notice that $L(\hat{b}_{\mathcal{D}})$ is *model-specific* and $\hat{b}_{\mathcal{D}}$ is random (Eq. (A5)). Taking expectation over $\hat{b}_{\mathcal{D}}$ leads to the *data-specific* generalization error

$$\begin{aligned}\mathbb{E}_{\hat{b}_{\mathcal{D}}}[L(\hat{b}_{\mathcal{D}})] &= \int_{\hat{b}_{\mathcal{D}}} \sum_{i=1}^{d_x+d_u} \Delta\hat{b}_{\mathcal{D},i}^2 p(\hat{b}_{\mathcal{D}}) d\hat{b}_{\mathcal{D}} \\ &= \sum_{i=1}^{d_x+d_u} \int_{\hat{b}_{\mathcal{D}}} \Delta\hat{b}_{\mathcal{D},i}^2 p(\hat{b}_{\mathcal{D},i}) p(\hat{b}_{\mathcal{D},-i}|\hat{b}_{\mathcal{D},i}) d\hat{b}_{\mathcal{D}} \\ &= \sum_{i=1}^{d_x+d_u} \int_{\hat{b}_{\mathcal{D},i}} \Delta\hat{b}_{\mathcal{D},i}^2 p(\hat{b}_{\mathcal{D},i}) d\hat{b}_{\mathcal{D},i} \\ &= \sum_{i=1}^{d_x+d_u} Var(\Delta\hat{b}_{\mathcal{D},i}) \\ &= (\sigma_y^2 + b_s^2\sigma_s^2) tr((D^T D)^{-1}).\end{aligned} \tag{A9}$$

Here $\hat{b}_{\mathcal{D},-i}$ are elements of $\hat{b}_{\mathcal{D}}$ other than the ith element. The derivation used the fact that

$$\int_{\hat{b}_{\mathcal{D},-i}} p(\hat{b}_{\mathcal{D},-i}|\hat{b}_{\mathcal{D},i}) d\hat{b}_{\mathcal{D},-i} = 1, \tag{A10}$$

and that $\mathbb{E}[\Delta\hat{b}_{\mathcal{D},i}] = 0$.

Finally, to derive the generalization error for the questionnaire (and the use of OLS), we need to compute

$$\mathbb{E}_{\mathcal{D},\varepsilon_y,w}[L(\hat{b}_{\mathcal{D}})] = \mathbb{E}_{\mathcal{D}}[(\sigma_y^2 + b_s^2\sigma_s^2) tr((D^T D)^{-1})].$$

Notice that from (R1) we have $D^T D = N_1 I \in \mathbb{R}^{(d_x+d_u)\times(d_x+d_u)}$. This leads to the solution

$$L_1 := \mathbb{E}_{\mathcal{D},\varepsilon_y,w}[L(\hat{b}_{\mathcal{D}})] = (\sigma_y^2 + b_s^2\sigma_s^2) \frac{d_x+d_u}{N_1}. \tag{A11}$$

The solution in Eq. (A11) is consistent with intuition: (1) Increasing noise in response (σ_y^2 and σ_s^2) increases the generalization error. In particular, the part-worth of the form score scales the influence of σ_s^2; and (2) The dimensionalities of s and u, and the sample size, influence the generalization error in opposite directions.

A3 Case 2: Bi-phase Questionnaire

A3.1 Data

To analyze the second case, we consider the existence of data $\mathcal{D} = \{y^i, x^i, u^i, s^i\}_{i=1}^{N_2}$, i.e., in addition to what we can observe in Case 1, *now we also observe the form score*. This is again a simplification from the actual experiments, where only comparisons between forms rather than their actual scores are observable. We adopt the same assumptions as in Case 1.

A3.2 Prediction and uncertainty

With the data $\mathcal{D}$, we can build two OLS models. For form score, we have the prediction

$$\hat{s}_{\mathcal{D}} = \hat{\beta}_{\mathcal{D}}^T x, \tag{A12}$$

and for utility

$$\hat{y}_{\mathcal{D}} = \hat{b}_{s_{\mathcal{D}}}^{T} \hat{\beta}_{\mathcal{D}}^{T} x + \hat{b}_{u_{\mathcal{D}}}^{T} u, \tag{A13}$$

where $\hat{\beta}_{\mathcal{D}}$, $\hat{b}_{s_{\mathcal{D}}}$, $\hat{b}_{u_{\mathcal{D}}}$ are OLS estimates of β, b_s, b_u, respectively. For simplicity, we introduce $\hat{\theta}_{\mathcal{D}} := \{\hat{\beta}_{\mathcal{D}}, \hat{b}_{s_{\mathcal{D}}}, \hat{b}_{u_{\mathcal{D}}}\}$.

According to our models, the variance-covariance matrix of $\hat{\beta}_{\mathcal{D}}$ is

$$Var(\hat{\beta}_{\mathcal{D}}) = \sigma_s^2 (X^T X)^{-1} = \frac{\sigma_s^2}{N_2} I. \tag{A14}$$

The variance-covariance matrix of $[\hat{b}_{s_{\mathcal{D}}}, \hat{b}_{u_{\mathcal{D}}}^{T}]$ is

$$Var([\hat{b}_{s_{\mathcal{D}}}, \hat{b}_{u_{\mathcal{D}}}^{T}]) = \sigma_y^2 ([s, U]]^T [s, U])^{-1}. \tag{A15}$$

Here $s := [s^1, \ldots, s^{N_2}]^T$ are the collected form scores.

A3.3 Generalization error

The model-specific generalization error is

$$\begin{aligned} L(\hat{\theta}_{\mathcal{D}}) &= \mathbb{E}_w \left[(\hat{b}_{s_{\mathcal{D}}} \hat{\beta}_{\mathcal{D}}^T x + \hat{b}_{u_{\mathcal{D}}}^{T} u - b_s \beta^T x - b_u^T u)^2 \right] \\ &= \mathbb{E}_w \left[((\hat{b}_{s_{\mathcal{D}}} \hat{\beta}_{\mathcal{D}} - b_s \beta)^T x + (\hat{b}_{u_{\mathcal{D}}} - b_u)^T u)^2 \right] \\ &= \sum_{i=1}^{d_x} (\Delta \hat{b}_{s_{\mathcal{D}}} \hat{\beta}_{\mathcal{D},i})^2 + \sum_{j=1}^{d_u} (\Delta \hat{b}_{u_{\mathcal{D},j}})^2. \end{aligned} \tag{A16}$$

Here $\Delta \hat{b}_{s_{\mathcal{D}}} \hat{\beta}_{\mathcal{D},i}$ is the ith element of $\hat{b}_{s_{\mathcal{D}}} \hat{\beta}_{\mathcal{D}} - b_s \beta$, and $\Delta \hat{b}_{u_{\mathcal{D},j}}$ is the jth element of $\hat{b}_{u_{\mathcal{D}}} - b_u$. Similar to Case 1, the derivation requires the assumptions (R1-3).

The data-specific generalization error is

$$\mathbb{E}_{\hat{\theta}_{\mathcal{D}}}[L(\hat{\theta}_{\mathcal{D}})] = \mathbb{E}_{\hat{\theta}_{\mathcal{D}}} \left[\sum_{i=1}^{d_x} (\Delta \hat{b}_{s_{\mathcal{D}}} \hat{\beta}_{\mathcal{D},i})^2 + \sum_{j=1}^{d_u} (\Delta \hat{b}_{u_{\mathcal{D},j}})^2 \right] \tag{A17}$$

We first compute the second term on the right-hand side (RHS):

$$\mathbb{E}_{\hat{b}_{u_{\mathcal{D}}}} \left[\sum_{j=1}^{d_u} (\Delta \hat{b}_{u_{\mathcal{D},j}})^2 \right] = \sum_{j=1}^{d_u} Var(\hat{b}_{u_{\mathcal{D},j}}) = tr(Var(\hat{b}_{u_{\mathcal{D}}})). \tag{A18}$$

From Eq. (A15) and using the analytical inverse of 2-by-2 block matrices, we have

$$Var(\hat{b}_{u_{\mathcal{D}}}) = \frac{\sigma_y^2}{N_2} \left(I + \frac{U^T s s^T U}{N_2 s^T s - s^T U U^T s} \right) \tag{A19}$$

Now we compute the first term on the RHS of Eq. (A17):

$$\begin{aligned}\mathbb{E}_{\hat{b}_{s_\mathcal{D}},\hat{\beta}_\mathcal{D}}\left[\sum_{i=1}^{d_x}(\Delta\hat{b}_{s_\mathcal{D}}\hat{\beta}_{\mathcal{D},i})^2\right] &= \mathbb{E}_{\hat{b}_{s_\mathcal{D}},\hat{\beta}_\mathcal{D}}\left[\sum_{i=1}^{d_x}\left((\hat{b}_{s_\mathcal{D}}\hat{\beta}_{\mathcal{D},i})^2+(b_s\beta)^2-2\hat{b}_{s_\mathcal{D}}\hat{\beta}_{\mathcal{D},i}b_s\beta\right)\right]\\ &= \sum_{i=1}^{d_x}\left(\mathbb{E}_{\hat{b}_{s_\mathcal{D}}}[\hat{b}_{s_\mathcal{D}}^2]\mathbb{E}_{\hat{\beta}_\mathcal{D}}[\hat{\beta}_\mathcal{D}^2]-(b_s\beta_i)^2\right)\\ &= \sum_{i=1}^{d_x}\left(\left(Var(\hat{b}_{s_\mathcal{D}})+b_s^2\right)\left(Var(\hat{\beta}_{\mathcal{D},i})+\beta_i^2\right)-(b_s\beta_i)^2\right)\\ &= \sum_{i=1}^{d_x}\left(Var(\hat{b}_{s_\mathcal{D}})Var(\hat{\beta}_{\mathcal{D},i})+b_s^2Var(\hat{\beta}_{\mathcal{D},i})+\beta_i^2Var(\hat{b}_{s_\mathcal{D}})\right)\\ &= tr(Var(\hat{\beta}_\mathcal{D}))\left(Var(\hat{b}_{s_\mathcal{D}})+b_s^2\right)+\beta^T\beta Var(\hat{b}_{s_\mathcal{D}}).\end{aligned} \tag{A20}$$

From Eq. (A15) and using the analytical inverse of 2-by-2 block matrices, we have

$$Var(\hat{b}_{s_\mathcal{D}}) = \sigma_y^2\left(s^Ts-\frac{s^TUU^Ts}{N_2}\right)^{-1} \tag{A21}$$

Lastly, we need to compute the generalization error for the questionnaire:

$$\mathbb{E}_{\mathcal{D},\varepsilon_s,\varepsilon_y,w}[L(\hat{\theta}_\mathcal{D})] = \mathbb{E}_\mathcal{D}\left[tr\left(Var(\hat{b}_{u_\mathcal{D}})\right)+tr\left(Var(\hat{\beta}_\mathcal{D})\right)\left(Var(\hat{b}_{s_\mathcal{D}})+b_s^2\right)+\beta^T\beta Var(\hat{b}_{s_\mathcal{D}})\right] \tag{A22}$$

In the following, we will derive the upper bound of $\mathbb{E}_{\mathcal{D},\varepsilon_s,\varepsilon_y,w}[L(\hat{\theta}_\mathcal{D})]$ through those of $\mathbb{E}_\mathcal{D}\left[Var(\hat{b}_{s_\mathcal{D}})\right]$ and $\mathbb{E}_\mathcal{D}\left[tr\left(Var(\hat{b}_{u_\mathcal{D}})\right)\right]$. We will then compare the resultant upper bound with the generalization error from Case 1.

A3.4 Upper bound of $\mathbb{E}_\mathcal{D}\left[Var(\hat{b}_{s_\mathcal{D}})\right]$

We start by studying

$$\mathbb{E}_\mathcal{D}\left[Var(\hat{b}_{s_\mathcal{D}})\right] = \sigma_y^2\mathbb{E}_\mathcal{D}\left[\left(s^Ts-\frac{s^TUU^Ts}{N_2}\right)^{-1}\right] \tag{A23}$$

Notice that

$$s^Ts-\frac{s^TUU^Ts}{N_2} = \beta^TX^TX\beta+\varepsilon_s^TX\beta+\varepsilon_s^T(I-\frac{UU^T}{N_2})\varepsilon_s \tag{A24}$$

Using (R1), we have $\beta^TX^TX\beta=\beta^T\beta$. For the second RHS term of Eq. (A24), we have $\varepsilon_s^TX\beta=\sum_{i=1}^{N_2}\varepsilon_s^i\beta^Tx^i$, where ε_s^i and x^i are the error of the ith sample, and the ith input, respectively. Recall that from the data model, $\varepsilon_s^i\overset{\text{i.i.d.}}{\sim}N(0,\sigma_s^2)$ and $x^i\overset{\text{i.i.d.}}{\sim}N(0,1)$. Therefore there exists $q>1$,

$$|\varepsilon_s^i\beta^Tx^i| < q\sigma_s||\beta||_1 \tag{A25}$$

with high probability, where $||\beta||_1$ is the l_1-norm of β. To make this argument more concrete, we can modify the data model to let ε_s^i and x^i follow truncated normal distributions, so that Eq. (A25) holds almost surely. For example, we can truncate $|\varepsilon_s^i|$ at $3\sigma_s$, and $|x^i|$ at 3, leading to $q=9$. With these, we have

$$\mathbb{E}_{X,\varepsilon_s}[\varepsilon_s^TX\beta] \geq -N_2q\sigma_s||\beta||_1 \tag{A26}$$

For the last term on the RHS of Eq. (A26), we first introduce Lemma 1.

Lemma 1 *Let $U \in \mathbb{R}^{N_2 \times d_u}$ be a random matrix where each element $u_{j,i}$ is i.i.d. standard normal. For sufficiently large N_2 and small d_u, $N_2 I - UU^T$ is p.d. (positive definite) with high probability. In particular, when $(N_2, d_u) = (100,2)$, the probability for $N_2 I - UU^T$ to not be p.d. is less than 0.01; for $(N_2, d_u) = (1000,2)$, the probability is less than 10^{-10}.*

Proof. First we note that

$$N_2 I - UU^T = \sum_{i=1}^{d_u} \left(\frac{N_2}{d_u} I - u_{\cdot,i} u_{\cdot,i}^T \right). \tag{A27}$$

We will show that $A_i = \frac{N_2}{d_u} I - u_{\cdot,i} u_{\cdot,i}^T$ is p.d. with high probability for all i. To this end, we inspect each row of A_i. For the jth row, the sum of its elements (denoted by $a_{i,j}$) is

$$a_{i,j} = \frac{N_2}{d_u} - u_{j,i}^2 - \sqrt{N_2 - 1} \frac{\sum_{k \neq j} u_{k,i}}{\sqrt{N_2 - 1}} u_{j,i}. \tag{A28}$$

Here $\frac{\sum_{k \neq j} u_{k,i}}{\sqrt{N_2 - 1}}$ and $u_{j,i}$ are i.i.d. standard normal. From Lemma 2 (below), we have

$$P\left(\frac{\sum_{k \neq j} u_{k,i}}{\sqrt{N_2 - 1}} u_{j,i} > t \right) \leq \exp\left(-h(t)t - \frac{1}{2} \ln(1 - h(t)^2) \right), \tag{A29}$$

where

$$h(t) = \frac{\sqrt{1 + 4t^2} - 1}{2t}. \tag{A30}$$

We also know that $u_{j,i}^2$ follows a chi-square distribution with 1 d.f. Therefore, we can compute the following probability

$$P(a_{i,j} \leq 0) = \int_0^\infty P\left(\frac{\sum_{k \neq j} u_{k,i}}{\sqrt{N_2 - 1}} u_{j,i} \geq \frac{N_2}{d_u} I - t \right) p(u_{j,i}^2 = t) dt. \tag{A31}$$

The above probability can be calculated numerically; e.g., for $(N_2, d_u) = (100,2)$, $P(a_{i,j} \leq 0) = 2.5 \times 10^{-5}$, and for $(N_2, d_u) = (1000,2)$, $P(a_{i,j} \leq 0) = 1.7 \times 10^{-14}$.

A sufficient condition for $N_2 I - UU^T$ to be p.d. is for A_i to be p.d. for all i. Since $a_{i,j}$ for all js are correlated, we have

$$P(a_{i,j} > 0, \forall i, j) \geq \prod_{i=1}^{d_u} \prod_{j=1}^{N_2} \left(1 - P(a_j \leq 0) \right) \approx 1 - d_u N_2 P(a_j \leq 0). \tag{A32}$$

For $(N_2, d_u) = (100,2)$, $1 - P(a_{i,j} > 0, \forall i, j) \leq 0.01$ and for $(N_2, d_u) = (1000,2)$, $1 - P(a_{i,j} > 0, \forall i, j) \leq 10^{-10}$.

Lemma 2 *Let X and Y be i.i.d. standard normal, and $Z = XY$. We have*

$$P(Z > t) \leq \exp\left(-h(t)t - \frac{1}{2} \ln(1 - h(t)^2) \right), \tag{A33}$$

where

$$h(t) = \frac{\sqrt{1+4t^2}-1}{2t}. \tag{A34}$$

Proof. We first note that the random set $\{X, Y\}$ is equal in distribution to the random set $\{(X - Y)/\sqrt{2}, (X + Y)/\sqrt{2}\}$. Therefore $Z = XY$ is equal in distribution to $(X^2 - Y^2)/2$. So the moment generating function of Z is

$$\begin{aligned} \mathbb{E}_Z[\exp(hZ)] &= \mathbb{E}_{X,Y}\left[\exp\left(\frac{h}{2}(X^2 - Y^2)\right)\right] \\ &= \mathbb{E}_X\left[\exp\left(\frac{h}{2}X^2\right)\right]\mathbb{E}_Y\left[\exp\left(-\frac{h}{2}Y^2\right)\right] \\ &= (\sqrt{1-h^2})^{-1} \end{aligned} \tag{A35}$$

Then we have

$$\begin{aligned} P(Z > t) &\leq \inf_{h\geq 0}\mathbb{E}_Z[\exp(h(Z-t))] \\ &= \inf_{h\geq 0}\exp\left(-ht - \frac{1}{2}\ln(1-h^2)\right). \end{aligned} \tag{A36}$$

Let $l(\theta) := -ht - \frac{1}{2}\ln(1-h^2)$, its minimum is reached when

$$h = \frac{\sqrt{1+4t^2}-1}{2t}. \tag{A37}$$

From Lemma 1, $\varepsilon_S^T(I - \frac{UU^T}{N_2})\varepsilon_S > 0$ with probability close to 1 for reasonably large N_2 and small d_u. By further truncating the distribution of u^i, we can make this statement certain: $\varepsilon_S^T(I - \frac{UU^T}{N_2})\varepsilon_S > 0$ almost surely for reasonably large N_2 and small d_u. We can now use Jensen's inequality and Eq. (A26) to get

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\left[Var(\hat{b}_{s_{\mathcal{D}}})\right] &\leq \sigma_y^2\mathbb{E}_{\mathcal{D}}[(\beta^T X^T X\beta + \varepsilon_S^T X\beta)^{-1}] \\ &\leq \sigma_y^2(\beta^T X^T X\beta - N_2 q\sigma_s||\beta||_1)^{-1} \\ &= \frac{\sigma_y^2}{N_2}(\beta^T\beta - q\sigma_s||\beta||_1)^{-1} \end{aligned} \tag{A38}$$

A3.5 Upper bound of $\mathbb{E}_{\mathcal{D}}\left[tr\left(Var(\bar{b}_{u_{\mathcal{D}}})\right)\right]$

In order to derive and upper bound, we now study

$$\mathbb{E}_{\mathcal{D}}\left[tr\left(Var(\bar{b}_{u_{\mathcal{D}}})\right)\right] = \mathbb{E}_{\mathcal{D}}\left[tr\left(\frac{\sigma_y^2}{N_2}\left(I + \frac{U^T ss^T U}{N_2 s^T s - s^T UU^T s}\right)\right)\right]. \tag{A39}$$

The trace on the RHS has the following upper bound

$$
\begin{aligned}
tr\left(\frac{\sigma_y^2}{N_2}\left(I+\frac{U^T s s^T U}{N_2 s^T s - s^T U U^T s}\right)\right) &= \frac{\sigma_y^2 d_u}{N_2} + \frac{\sigma_y^2}{N_2}\frac{tr(U^T s s^T U)}{N_2 s^T s - s^T U U^T s} \\
&\leq \frac{\sigma_y^2 d_u}{N_2} + \frac{\sigma_y^2 tr(U^T s s^T U)}{N_2^3(\beta^T\beta - q\sigma_s||\beta||_1)}.
\end{aligned} \tag{A40}
$$

We also have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[tr(U^T s s^T U)] &= \mathbb{E}_{\mathcal{D}}[tr(U^T(X\beta+\varepsilon_s)(X\beta+\varepsilon_s)^T U)] \\
&= \mathbb{E}_{\mathcal{D}}[tr(U^T \varepsilon_s \varepsilon_s^T U)] \\
&= \mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{d_u}\left(\sum_{k=1}^{N_2} u_i^k \varepsilon_s^k\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{d_u}\left(\sum_{k=1}^{N_2}\left(u_i^k \varepsilon_s^k\right)^2 + 2\sum_{k=1}^{N_2}\sum_{k'>k}\left(u_i^k \varepsilon_s^k u_i^{k'} \varepsilon_s^{k'}\right)\right)\right] \\
&= \sum_{i=1}^{d_u}\left(\sum_{k=1}^{N_2}\mathbb{E}_{\mathcal{D}}\left[\left(u_i^k \varepsilon_s^k\right)^2\right] + 2\sum_{k=1}^{N_2}\sum_{k'>k}\mathbb{E}_{\mathcal{D}}\left[u_i^k \varepsilon_s^k u_i^{k'} \varepsilon_s^{k'}\right]\right) \\
&= d_u N_2 \sigma_s^2.
\end{aligned} \tag{A41}
$$

Plug Eq. (A40) and Eq. (A41) into Eq. (A39) to get

$$
\mathbb{E}_{\mathcal{D}}\left[tr\left(Var(\hat{b}_{u_{\mathcal{D}}})\right)\right] \leq \frac{\sigma_y^2 d_u}{N_2}\left(1+\frac{\sigma_s^2}{N_2(\beta^T\beta - q\sigma_s||\beta||_1)}\right) \tag{A42}
$$

A3.6 Upper bound of the generalization error

Using results from Eq. (A14), Eq. (A38), and Eq. (A42), we have

$$
\mathbb{E}_{\mathcal{D},\varepsilon_s,\varepsilon_y,w}[L(\hat{\theta}_{\mathcal{D}})] \leq \frac{\sigma_y^2 d_u}{N_2}\left(1+\frac{\sigma_s^2}{N_2\gamma}\right) + \frac{d_x\sigma_s^2}{N_2}\left(\frac{\sigma_y^2}{N_2\gamma} + b_s^2\right) + \beta^T\beta\frac{\sigma_y^2}{N_2\gamma}, \tag{A43}
$$

where $\gamma = \beta^T\beta - q\sigma_s||\beta||_1$). Denote the upper bound as $\bar{L}_2 := \mathbb{E}_{\mathcal{D},\varepsilon_s,\varepsilon_y,w}[L(\hat{\theta}_{\mathcal{D}})]$. When N_2 is a large number, the upper bound can be approximated as

$$
\bar{L}_2 \approx \frac{\sigma_y^2}{N_2}\left(d_u + \frac{\beta^T\beta}{\gamma}\right) + \frac{b_s^2 d_x \sigma_s^2}{N_2} \tag{A44}
$$

A4 Comparison between $\bar{L}_2$ and L_1

Using results from Eq. (A11) and Eq. (A44), and denoting $N_2/N_1 = \alpha$, we can now derive the lower bound of the difference between generalization errors from Case 1 and Case 2:

$$
L_1 - \bar{L}_2 = \left(d_x - \frac{\beta^T\beta}{\alpha\gamma} + d_u\left(1-\frac{1}{\alpha}\right)\right)\frac{\sigma_y^2}{N_1} + \left(d_x + d_u - \frac{d_x}{\alpha}\right)\frac{\sigma_s^2}{N_1}. \tag{A45}
$$

A few remarks can be made from Eq. (A45), as follows.

A4.1 Remark 1

When the form responses are noiseless, i.e., $\sigma_s^2 = 0$, $L_1 - \bar{L}_2 \geq 0$ if $d_x \geq 1/\alpha$. Consider the specific case where $N_1 = 2N_2$ ($\alpha = 0.5$). Then $d_x \geq 2$ will suffice for questionnaire 2 to have generalization error no greater than questionnaire 1.

A4.2 Remark 2

When $\sigma_s^2 \neq 0$, $L_1 - \bar{L}_2 \geq 0$ if the form signal-to-noise ratio $||\beta||_1/\sigma_s$ is sufficiently large, and the form-to-utility noise ratio σ_s^2/σ_y^2 is sufficiently small. To show this, we start by rearranging Eq. (A45) to show that $L_1 - \bar{L}_2 \geq 0$ if

$$\sigma_s^2 \leq \left(\frac{\alpha - \frac{||\beta||_2^2}{\gamma d_x}}{1-\alpha} \right) \sigma_y^2. \tag{A46}$$

Because σ_s^2 is positive, the above condition cannot hold when $\alpha\gamma d_x < ||\beta||_2^2$. Using the fact that

$$||\beta||_2^2 \leq ||\beta||_1^2, \tag{A47}$$

we can derive the following sufficient condition when $L_1 - \bar{L}_2 < 0$:

$$\begin{aligned} & \alpha\gamma d_x < ||\beta||_2^2 \\ \Rightarrow \quad & \alpha(||\beta||_2^2 - q\sigma_s||\beta||_1)d_x < ||\beta||_2^2 \\ \Rightarrow \quad & \alpha q\sigma_s||\beta||_1 d_x > (\alpha d_x - 1)||\beta||_2^2 \\ \Rightarrow \quad & \alpha q\sigma_s d_x > (\alpha d_x - 1)||\beta||_1 \\ \Rightarrow \quad & ||\beta||_1 < \frac{\alpha q\sigma_s d_x}{\alpha d_x - 1}. \end{aligned} \tag{A48}$$

For the specific case where d_x is large and $\alpha = 0.5$, the above condition can be reduced to

$$||\beta||_1 < q\sigma_s \tag{A49}$$

■

APPENDIX B (Online): Accommodating the "No Choice" Option

Because our purchase question can be viewed as a forced binary choice, we explored the changes required to accommodate a "no choice" option as well. This can be done via a minor extension of Eq. (3) and (4). The primal problem becomes:

$$\min_{\mathbf{w}} \mathbf{w}^T\mathbf{w}$$
$$\text{subject to } \mathbf{w}^T\phi\left(\mathbf{x}_j^{(1)}\right) - \mathbf{w}^T\phi\left(\mathbf{x}_j^{(2)}\right) \geq c_j, \forall \text{ choices } j \text{ other than no choice}, c_j \in \{1,2\}$$
$$1 \geq \mathbf{w}^T\phi\left(\mathbf{x}_{j'}^{(1)}\right) - \mathbf{w}^T\phi\left(\mathbf{x}_{j'}^{(2)}\right) \geq -1, \forall \text{ choices } j' \text{as "no choice"}$$

The dual is then:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - [\mathbf{c}^T, \mathbf{1}^T, -\mathbf{1}^T]\boldsymbol{\alpha}$$
$$\text{subject to: } \boldsymbol{\alpha} \geq \mathbf{0}.$$

The additional constants in the linear term of the objective come from the constraints related to the "no choice" data. Since this problem has exactly the same formulation as the one examined in the paper, the same algorithm applies.

APPENDIX C (Online): Lagrangian Formulation of Eq. (3)

We note again that the SVM formulation does not explicitly rely on a notion of likelihood, Yet it is possible to treat the convex SVM objective as a negative log-likelihood using hinge-loss and a simple i.i.d. Gaussian prior; see, for example, Evgeniou, Pontil, and Toubia (2007).

For the problem formulation in (3), first note that its Lagrangian is:

$$L(\mathbf{w}, \boldsymbol{\alpha}) = \mathbf{w}^T\mathbf{w} + \sum_j \alpha_j \left(c_j - \mathbf{w}^T \left(\boldsymbol{\phi}\left(\mathbf{x}_j^{(1)}\right) - \boldsymbol{\phi}\left(\mathbf{x}_j^{(2)}\right) \right) \right).$$

$L(\mathbf{w}, \boldsymbol{\alpha})$ can be considered as a negative log-likelihood with parameters $\boldsymbol{\alpha}$ balancing the Gaussian prior ($\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$) and the data ($\mathbf{x}_j^{(1)}$ more preferred than $\mathbf{x}_j^{(2)}$). The likelihood has the following form:

$$L(\mathbf{w}; \{(\mathbf{x}_1, \mathbf{x}_2)_i, c_i\}) = \exp(-\frac{1}{2}\mathbf{w}^T\mathbf{w}) \prod_j \exp\left(-\frac{\alpha_j}{2} \left(c_j - \mathbf{w}^T \left(\boldsymbol{\phi}\left(\mathbf{x}_j^{(1)}\right) - \boldsymbol{\phi}\left(\mathbf{x}_j^{(2)}\right) \right) \right) \right),$$

which can be further simplified as

$$L(\mathbf{w}; \{(\mathbf{x}_1, \mathbf{x}_2)_i, c_i\}) = \exp(-\frac{1}{2}\mathbf{w}^T\mathbf{w}) \prod_j \exp\left(-\frac{\alpha_j}{2} \left(c_j - \mathbf{w}^T\mathbf{z}_j \right) \right),$$

with $\mathbf{w}$ the estimated parameters, $\mathbf{z}_j = \boldsymbol{\phi}\left(\mathbf{x}_j^{(1)}\right) - \boldsymbol{\phi}\left(\mathbf{x}_j^{(2)}\right)$ the data, and $\boldsymbol{\alpha}$ hyperparameters. This indicates that the probability of choosing $\mathbf{x}_j^{(1)}$ over $\mathbf{x}_j^{(2)}$ follows the PMF:

$$\Pr\left(\mathbf{x}_j^{(2)} \succ \mathbf{x}_j^{(2)}; \mathbf{w}\right) = \begin{cases} 1, \mathbf{w}^T\mathbf{z}_j \geq c_j \\ \exp\left(-\frac{\alpha_j}{2} \left(c_j - \mathbf{w}^T\mathbf{z}_j \right) \right), \text{otherwise} \end{cases}.$$

Several differences between our model and Evgeniou et al.'s (2007) should be mentioned:

(1) We use a Gaussian kernel to define $\boldsymbol{\phi}^T\boldsymbol{\phi}$

(2) Due to the infinite dimension of the feature space induced by the Gaussian kernel, we do not model or learn the variance-covariance matrix of the part-worths. This matrix is assumed to be identity (as is indicated by the shrinkage term).

(3) We treat the data as constraints in the form: $\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) \geq c$ for all $\mathbf{x}_1$ preferred to $\mathbf{x}_2$, while in Evgeniou et al. (2007), data are used to form a quadratic loss in the objective: $\left(1 - \mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2)\right)^2$, for which an analytical solution exists.

[1] Evgeniou, Theodoros, Massimiliano Pontil, and Olivier Toubia. "A convex optimization approach to modeling consumer heterogeneity in conjoint estimation." *Marketing Science* 26, no. 6 (2007): 805-818.

APPENDIX D (Online): Insensitivity of Individual-Level Learning to Constraint Value

The constant in Eq. (10) is set exogenously to 1. Note that, when it's set to 0, a trivial solution ($w = 0$) becomes the optimal one. To prevent this from happening, the hard-margin SVM formulation uses "1" as the minimal gap between the preferred and non-preferred choices.

The following explains why the constant gap can be set arbitrarily. First, similar to the conversion from Eq. (3) to Eq. (4), here the dual of Eq. (10) is:

$$\text{(P1)} \quad \min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha}$$
$$\text{subject to: } \boldsymbol{\alpha} \geq \mathbf{0},$$

where the coefficient vector ($\mathbf{1}$) of the linear term in the objective is contributed by the constant gap in the primal. Let its solution be $\boldsymbol{\alpha}_1^*$, and now also let the gap be c. The resulting dual becomes:

$$\text{(P2)} \quad \min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - c\mathbf{1}^T\boldsymbol{\alpha}$$
$$\text{subject to: } \boldsymbol{\alpha} \geq \mathbf{0},$$

It is readily seen that the solution of (P2) is $c\boldsymbol{\alpha}_1^*$, and thus the solution to the primal of P2 will also be a scaled version of that of Eq. (10). **This means that the choice of the gap will not affect the preference ranking**, i.e., if one product is preferred to the other under the model derived from (P1), the same is true under that from (P2).

We note, however, that when two gaps are introduced, as is the case in Eq. (4), the choice of the gap values (in the paper, these are 1 and 2) does affect the solution, i.e., choosing other numbers than 2 may lead to a preference model with different rankings. We examine this case in detail in Appendix G.

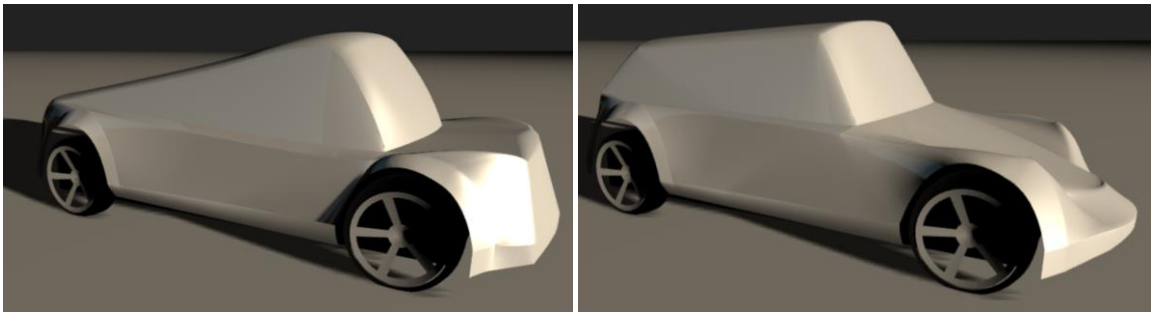
APPENDIX E (Online): Setting Weights (v_1 and v_2) in Eq. 16

As noted in the paper proper, $v_1 = 0.99$ and $v_2 = 0.01$ were selected for the experiments. Here we explore why, as well as the effects of alternative settings.

First, note that the lopsided weighting ($v_1/v_2 = 99$) is a result of the finding that when v_2 is large, the generated design vector (x) tends to be populated by zeros and ones. This is reasonable, since such vectors are the corners of the 19-dimensional design space, and so are truly "apart" from one another. This is consistent with the curse of dimensionality: higher-dimensional Euclidean space has larger corner volumes.

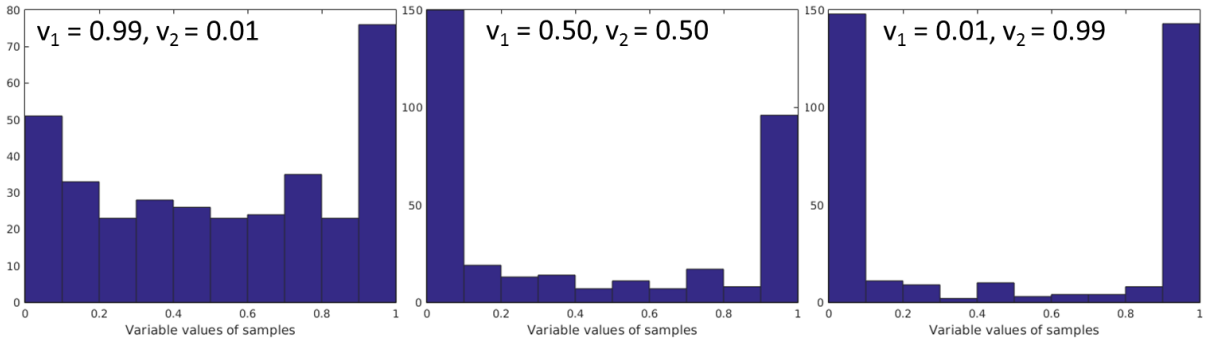
Based on this finding, we empirically tested various settings of the weights. The chosen ones produce designs off the boundaries of the design space, and also achieve reasonable hit-rate on the validation set during an internal preliminary test. [Specifically, we conducted many pilot surveys before the final one, and tuned parameters through those pilot surveys.]

While these settings do satisfy the need to explore the design space, the resultant user comparisons between extreme designs may not lead to effective learning of preference models that can distinguish designs with mild parametric changes. In addition, we hypothesize (but cannot validate in the absence of attitudinal data) that users will be less involved and feel confused when asked to nearly always compare designs with unusual appearances. Here is one such example:



For this reason, we would like the samples to have less extreme values, which motivates the presented settings of $v_1 = 0.99$ and $v_2 = 0.01$.

More concretely, here are three distributions of the sample $\mathbf{x}_2^{new}$ with v_1 set to 0.99 (the value used), 0.50 (equal weighting), and 0.01 (reversed weighting):



These results help justify the empirical choice of $v_1 = 0.99$ and $v_2 = 0.01$. This is because only the first provides a reasonably uniform distribution over the variable values, while even equal weighting is heavily U-shaped.

[**Details on producing these figures**: The style function follows the first simulation scenarios in the paper. The GA algorithm follows the same settings as in the user experiments: We run 100 generations with population size 20 for Eq. (15), and 500 generations with population size 50 for Eq. (16). We draw one pair of samples from $[0,1]^{19}$ initially, and sample the space by solving Eq. (15) and Eq. (16) for 18 iterations while accumulating the samples. The plotted distributions combine values from the 18 samples of $\mathbf{x}_2^{new}$.]

APPENDIX F (Online): Simultaneous vs. Sequential Optimization for Eq. (15) and (16)

Because the problems in Eq. (15) and Eq. (16) are highly nonconvex, they were solved using Genetic Algorithms (GAs); and, due to limited response time by the application engine, search terminated when the "max" number of iterations was reached. This "max" was in fact empirically tuned for the server we used (Google App engine), so that we could obtain responses within the server response limit (30 seconds) without causing glitches or long delays during user interactions.

Under this setting, one would still be able to obtain a response if one solved the following all-in-one problem instead of decomposing it into Eq. (15) and Eq. (16):

$$\max_{\mathbf{x}_1^{new},\mathbf{x}_2^{new}\in[0,1]^{19}} F_{aio}(\mathbf{x}_1^{new},\mathbf{x}_2^{new}) = v_1\exp(-\|S(\mathbf{x}_1^{new})-S(\mathbf{x}_2^{new})\|^2) + v_2(\|\mathbf{x}_1^{new}-\mathbf{x}_2^{new}\|^2 + \min_j\left\|\mathbf{x}_1^{new}-\mathbf{x}_j^{old}\right\|^2 + \min_j\left\|\mathbf{x}_2^{new}-\mathbf{x}_j^{old}\right\|^2)$$

We can compare the "solution qualities", i.e., the values of $F_{aio}(\mathbf{x}_1^{new},\mathbf{x}_2^{new})$ by solutions from solving the all-in-one problem directly and through Eq. (15) and Eq. (16) sequentially. The operative hypothesis is that the latter will tend to yield better solutions; because the problem is decomposable, searching in two smaller spaces is more effective than in the combined space, **given the same budget for searching**.

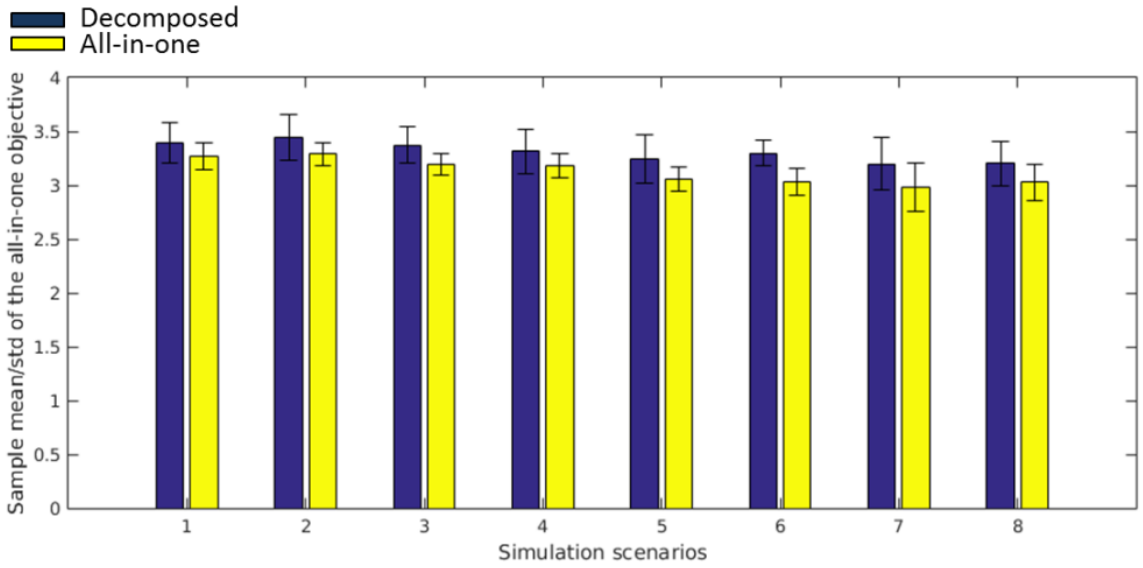
To test this hypothesis empirically, we conducted the following experiment. The style function follows the 8 simulation scenarios of Table 7:

Table 7: Consumer Preference Scenarios

Form importance	Response accuracy	Respondent hetero-geneity	Form score weight (λ)*	Form attribute coefficients: Independent terms (γ)	Form attribute coefficients: Interaction terms (δ)	Functional attribute partworths
Low	Low	Low	0.0043	N(0.5, 0.25)	N(0, 4.80)	N(0.5, 0.25)
Low	Low	High	0.0044	N(0.5, 1.5)	N(0, 13.7)	N(0.5, 1.5)
Low	High	Low	0.0028	N(3.0, 1.5)	N(0, 56.3)	N(3, 1.5)
Low	High	High	0.0057	N(3.0, 9.0)	N(0, 88.4)	N(3, 9.0)
High	Low	Low	0.0173	N(0.5, 0.25)	N(0, 4.80)	N(0.5, 0.25)
High	Low	High	0.0176	N(0.5, 1.5)	N(0, 13.7)	N(0.5, 1.5)
High	High	Low	0.0112	N(3.0, 1.5)	N(0, 56.3)	N(3.0, 1.5)
High	High	High	0.0230	N(3.0, 9.0)	N(0, 88.4)	N(3.0, 9.0)

Specifically, v_1 and v_2 are set to 0.99 and 0.01, respectively. The GA algorithm follows the same settings as in the user experiments: We run 100 generations with population size 20 for Eq. (15), and 500 generations with population size 50 for Eq. (16). For the all-in-one problem, we run 600 generations with population size 50. Note that this setting is slightly in favor of the all-in-one case as it uses the larger population size. We also draw 18 pairs of $\mathbf{x}^{old}$ uniformly from $[0,1]^{19}$ in order to calculate F_{aio}.

With these settings, we report a comparison between the F_{aio} values derived from the two optimization routines:



That the "decomposed" means have uniformly larger objective values provides substantial empirical support for the hypothesis and the pragmatic choice of using Eq. (15) and Eq. (16).

APPENDIX G (Online): Robustness Checks for Preference Gap Cutoffs, c_j, in Eq. (3)

In Eq. (3), c_j, was set to 1 and 2, which can entail substantive effects in the final solution. Here, we examine sensitivity to these choices. The sensitivity of the gap (1 for "better" and 2 for "much better") can be found from the Lagrange multipliers of the solution of the dual problem in Eq. (4). Investigating revealed that all Lagrange multipliers are positive (sample mean = 1.51, sample std = 0.81), i.e., all constraints in the SVM formulation are active. This indicates that the choice of the gaps (1 and 2) can indeed affect the styling preference model. This result, however, can be expected since only a relatively small number of questions are used to form each individual-level preference model in a high-dimensional space.

Note that the choice of the scale of the gap should not affect results substantially, since style preference is **further scaled by a part-worth** when forming the overall purchase preference, i.e., increasing the gap from 2 and decreasing the part-worth for styling will *maintain the landscape of the purchase preference with respect to shape-related variables*. The effects of the Gaussian parameter and the shrinkage parameters on generalization performance usually requires cross-validation, which may not be feasible for real-time interaction. Therefore, we set these parameters according to the default values used in the standard LibSVM package (e.g., www.csie.ntu.edu.tw/~cjlin/libsvm/).

We tested different c_j in Eq (3) for the 'worst' scenario in Table 8: high form importance, low response accuracy, and high heterogeneity. In this simulation, although {100, 200} yields the highest hit rate, the overall performance is relatively unaffected. We present these in two tables: {ratio fixed at ½; scale varies from 0.1 to 1000} and {ratios vary from ½ to 1/10, lower scale point fixed at unity}

<Choosing different c_j in Eq. (3)>

c_j	{0.1,0.2}	{1,2} (base)	{10,20}	{100,200}	{1000,2000}
Form preference hit rate	64.2	65.1	65.5	66	64.5

c_j	{1,2} (base)	{1,3}	{1,5}	{1,7}	{1,10}
Form preference hit rate	65.1	65.1	64.2	64.9	64.6

These values are nearly entirely invariant across a wide range of ratios and magnitudes, suggesting that the final results are not sensitive to this choice, at least for this problem. For others, manual tuning may be inevitable.

We note in closing that it is possible to introduce slack variables to form a soft SVM, which allows the gaps to be flexible rather than fixed to 1 and 2. However, as in any soft SVM formulation, doing so will inevitably introduce a hyperparameter that determines the balance between the training error (the sum of slacks) and the shrinkage. A typical way to tune the hyperparameter is through cross-validation. Thus, doing so during the online survey will roughly increase response time by a factor of N, where N is the number of folds used in cross-validation.

APPENDIX H (Online): Details on Query Engine and Response Times

Here we provide benchmarks regarding runtimes, since rapid query generation is critical for the success of any crowdsourced online survey platform. The live survey is hosted on a paid Google Application Engine site: https://vehiclechoicemodel.appspot.com.

We collected wall-clock response times for queries of the web application, using a lower- and a higher-end instance classes provided by Google App Engine. From the data, we conclude that (1) substantial differences exist between the response times for the two instance classes; and (2) later queries in the survey require longer response times. **Results show that scaling of the presented web app is feasible using existing cloud computing platforms in the market (e.g., Google App Engine).**

Details of the setup: The web application is implemented in JAVA 8 on the server side and javascript on the client side, and deployed to a Standard environment on Google App Engine. Two instance classes for comparison are the first-generation F1 and F4 (for JAVA 8, second-generation instances are not available). The specifications and costs of these instance classes are provided below [spec, price].

Instance class	Memory Limit	CPU Limit	Supported Scaling Types	Cost per hour per instance
F1	128 MB	600 MHz	Automatic	$0.05
F4	512 MB	2.4 GHz	Automatic	$0.20

We adopted the standard environment, but have not tested the response time of the app in a **flexible** environment, which is suitable for consistent traffic [doc]. The source code of the web app is available here [source; redacted for review], and the latest version of the web app can be accessed here [app].

Response time comparison: For each instance class, we record the response time for four *rounds* of queries. Each round consists of the following sequence of queries: Form ("F"), Attribute ("A"), Form and attribute ("F + A"), Form only. Except for the first "F" query which has pre-defined form parameters, each "F" query activates preference learning and the search of a new pair of forms. Each "A" query activates the search of a new pair of attributes for the given forms. Each "F + A" query activates the search of a new pair of forms and attributes. The "Form only" query only sends user choice data back to the server, and does not require any response from the server. Therefore, it has negligible response time.

Since the randomness in response time is insignificant, we conducted three independent surveys for each instance class. The mean and standard deviations of the response time (unit: sec.) are presented below:

F1	Round 1	Round 2	Round 3	Round 4
F + A	4.63 (1.57)	6.35 (0.23)	09.43 (0.35)	11.89 (0.10)
F	5.20 (0.43)	7.98 (0.54)	11.17 (0.42)	14.03 (0.34)
A	0.85 (0.26)	0.95 (0.23)	00.89 (0.08)	00.91 (0.06)
F4				
F + A	1.29 (0.06)	2.79 (0.04)	3.99 (0.14)	4.93 (0.11)
F	2.21 (0.16)	3.42 (0.09)	4.68 (0.19)	5.53 (0.13)
A	0.57 (0.11)	0.84 (0.08)	1.06 (0.24)	0.93 (0.06)

The results show that F4 processes queries effectively faster than F1, as expected, in particular for the heavier computations involved in "F + A" and "F" queries. We consider the average of these – roughly on the order of 3 seconds, and less than half the values for F1 –sufficient for commercial applications, and of course they can be further reduced with more computational power.

As the survey plays out, response times increase. This is because later searches have higher overhead, i.e., they incorporate samples from Form and Preference models learned based on previous rounds of survey questions. Lastly, the "F" queries take a slightly longer time than the "F + A" queries simply because the former are executed after the latter, and thus have slightly higher overhead in computation.

[spec] Google, The App Engine Standard Environment. URL: https://cloud.google.com/appengine/docs/standard/#instance_classes

[price] Google, App Engine Pricing. URL: https://cloud.google.com/appengine/pricing

[doc] Google, Choosing an App Engine environment. URL: https://cloud.google.com/appengine/docs/the-appengine-environments

[source] Source code for the paper. URL: https://github.com/RedactedForReview

[app] Web app for the paper. URL: https://vehiclechoicemodel.appspot.com/