Proceedings of the ASME 2023
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC/CIE 2023
August 20-23, 2023, Boston, Massachusetts

DETC2023-111714

ADAPTIVE OPTIMIZATION OF SUBJECTIVE DESIGN ATTRIBUTES: CHARACTERIZING INDIVIDUAL AND AGGREGATE PERCEPTIONS

Ananya Nandy
Dept. of Mechanical Engineering
University of California, Berkeley
Berkeley, CA, USA
ananyan@berkeley.edu

Kosa Goucher-Lambert
Dept. of Mechanical Engineering
University of California, Berkeley
Berkeley, CA, USA
kosa@berkeley.edu

ABSTRACT

Subjective attributes play a significant part in the assessment of user-facing products. Unlike performance requirements, these quantities are best evaluated through human feedback. While people share commonalities in their evaluations, allowing personalization when quantifying these subjective attributes may improve the alignment between computational and human representations of design information. We investigate this topic through a study in which participants (N = 56) make a series of pairwise decisions between parameterized mugs, and indicate their perceptions of how comfortable each is to hold. Interactive Bayesian optimization is used to adaptively arrive at a design that optimizes this subjective quantity. Participants guide the model through only their own decisions or make decisions using a model that has already been trained with simulated data (N = 25) or data from the real decisions of other participants (N = 31). The resulting designs are evaluated across the different cases, showing the impact of capturing individual and aggregate perceptions of subjective quantities. The findings imply that balancing aggregate and individual-level decisions within models simultaneously results in the best alignment with human perceptions of subjective attributes. Further implications for design include the potential for personalized control over subjective attributes for designers, users, or users-as-designers.

1. INTRODUCTION

In addition to satisfying functional requirements, product development relies on the creation of products that will be desirable to users in ways that are difficult to evaluate. While computational tools have the potential to augment designers' abilities, embedding human input directly within computational design approaches, leveraging the advantages of both humans and computers, is necessary to address subjective dimensions of design. For example, data regarding the higher-level use or context-related perceptions of products, referred to as product semantics [1], are highly subjective and difficult to represent computationally. Approaches to assess product semantics [2], or quantify user preferences more generally, make it possible to include human input into computational design methods [3–9]. A limitation of perceptual embeddings and many preference models is that they often represent group-level perceptions. However, individuals' perceptions may be at odds with this group-level perception. Prior work in design research has noted that significant differences can arise in preference from person to person [7]. Interest has increased in capturing "personal style" in the context of design [10]. In order to enable better interaction with designs at a subjective, semantic level, it may be helpful to individualize representations of these subjective attributes. In particular, the aim of the study is to gain insight into how individual differences may impact any quantification of subjective attributes in comparison to aggregate representations.

Focusing on the product semantic of "elegance," Poirson

et al. use interactive evolutionary computing (IEC) to move towards incorporating individualized perceptions [11]. While IEC can be used to find outcomes optimized for the subjective attribute, it can be helpful to learn a "function" that represents the perception of the attribute (much like a reward function in human-robot interaction [12]). Such a function can then be utilized for further decisions, such optimizing that attribute for a different design with the same high-level features or considering tradeoffs with other subjective attributes. An alternative method that has become popular for its flexibility and viability with smaller amounts of data is Bayesian optimization (BO). BO can reach an optimal outcome much like IEC, but importantly, it does so through a surrogate function that approximates the hard-to-evaluate, unknown function [13, 14] Using BO for individual users is particularly promising, for example, in the case of assistive technology such as exoskeleton gaits or hearing aids [15, 16]. This study builds upon prior work by applying a human-in-the-loop BO method (based on pairwise decisions and design modifications) to assess a subjective characteristic and empirically evaluate the alignment between human perceptions and the computationally optimized results. Pairwise decisions are collected from participants to model their perceptions of a subjective attribute through interactive Bayesian optimization, utilizing the test example of a parameterized drinking mug and how comfortable the mug will be to hold in their hands, individually. Highly individualized models are compared to those that involve aggregate data to assess if and how subtle individual perceptions can impact models of hard-to-measure quantities. Thus, a two-stage online study is used to address the following research question:

How do individual and aggregate models impact alignment of outcomes with a subjective attribute?

2. RELATED WORK

Prior research on methods for quantifying subjective attributes and preference modeling, areas that have been well studied within engineering design, are reviewed in this section. Relevant to the methods used here, Bayesian optimization and its applications in design or other domains are also reviewed.

2.1 Assessing Subjective Attributes for Design

There have been many efforts to quantify the perceptual space of user needs ("product semantics"). Within computer graphics, several approaches have been taken to map subjective semantics to 3D geometries. For example, geometric elements that preserve stylistic similarity between 3D shapes have been used for transferring styles to functionally compatible shapes [17]. Another approach has used crowdsourced pairwise comparisons to map subjective attributes (e.g. comfortable, sporty, etc.) directly to geometry using continuous deformation shape editing [18]. Several approaches have also been taken to model preferences related to product semantics so they can be integrated within design processes. Kansei engineering is a popular approach to extracting the desired emotion from a product [19]. Rating-based semantic differentials, such as those used in Kansei engineering, are one way to quantify subjective attributes and related preference. For instance, Reid et al. use ratings to quantify a specific product semantic, perceived environmental friendliness (PEF), generating new designs that better satisfy the subjective attribute [20]. This work also uses a singular subjective attribute, comfort, as an example. However, as ratings (or methods such as ranking or clustering) can require higher effort [21], pairwise decisions are used, as in [18]. Another approach to capturing product semantics has been to utilize multidimensional scaling to build similarity-based perceptual embedding spaces and relate this space to vectors of various semantic attributes [2,22,23]. Our prior work has also used perceptual embedding spaces for difficult-to-capture quantities [24]. However, a challenge is that these perceptual embeddings do not capture individual differences in how people make their judgments, which has been found to impact psychological spaces [25]. *Therefore, considering individualization is central to this research.* A simulated experiment also shows that when crowd-level preferences form, heuristics from the crowd information can increase the efficiency of eliciting preferences [26]. This is taken into account in this study by including a case where information from the "crowd" is used to initialize the model before individual decisions.

2.2 Preference Learning for Design

In this work, "preference" is considered only along a specific subjective dimension. However, more broadly, preference modeling has been applied to engineering design extensively. Early work in using preference learning techniques for product design uses a lottery question-based framework to create utility functions that reflect a designer's priorities [27]. In subsequent work, utility functions are considered extensively. Though the use of utility analysis has its limitations in engineering design, a major benefit is its ability to "model subjective tradeoffs, particularly those that are nonlinear and/or that must be made under uncertainty," which can be of particular use for adapting to individuals [28].

A common method to model preference is to determine the expected form of a utility function and then estimate weights via a discrete choice experiment, where participants make decisions between a number of choices (often pairwise). Approaches have been developed to incorporate form (generally, aesthetics) into these utility functions [3, 8, 29]. Often, preference models are used to analyze tradeoffs between function and a subjective attribute (most often form) [4, 7, 30]. The methods used in the above studies allow the analysis of attribute weightings to determine their impacts separately, but the presence of interaction

effects can be a challenge [31].

Prior work has also considered methods that can better capture nonlinearities inherent to subjective evaluation. For example, support vector machines (SVMs), Markov chains, genetic algorithms, or artificial neural networks (ANNs) have been used to map subjective attributes to design variables non-parametrically [5, 11, 32, 33]. Burnap et al. successfully demonstrate feature learning to predicting preferred designs. Feature learning is a promising approach to quantifying subjective semantic attributes at an individual level, but requires the collection of large amounts of user demographic data [34].

2.3 Active Querying and Bayesian Optimization

In many studies of preference, the choices to be presented to participants are determined ahead of time based on a random sample, D-efficient main effects, or Latin square design among others [6, 7, 20, 31]. However, active learning can address some of the challenges associated with design of experiments and individual differences. Active querying has been used for preference elicitation in engineering design, allowing quick convergence to a true utility function for a multi-objective problem [35]. Active preference learning has also been applied to finding the best product concepts when systematically assigning weights to product attributes is difficult [36]. Specifically, SVMs and ANNs have been used with a small number of active queries to find rankings for concepts that align with what an experienced designer selects manually. Active learning has also been applied to quantify form and function tradeoffs [9]. Although active learning is a challenge itself due to the high dimensionality of design spaces, this study uses Bayesian optimization and active learning in order to allow adaptive data collection. Bayesian optimization (closely related to Kriging models from geostatistics [37]) is a method that allows a blackbox function to be optimized. Within engineering design, Bayesian optimization has been used in cases when high-fidelity simulations are computationally expensive to run [38,39]. More importantly, it is particularly useful for the evaluation of subjective attributes since it is difficult to assume a functional form that will be appropriate for a person's judgments. Specifically, using Gaussian Processes (GPs) allows non-parametric estimation of a person's utility function, where the form of the function does not have to be specified ahead of time (but smoother functions are preferred) [13]. Active learning allows this to be done efficiently with as few evaluations of the quantity of interest as possible. Bayesian optimization methods have been explored in many domains, including visual parametric design, to tackle target-oriented cases when high-level feedback is easier to provide than tweaking parameters [40–42]. It has also been used in human-robot interaction to learn reward functions [12] and for exoskeleton gait optimization, using human feedback to find preferred gaits [15]. Relevant work within engineering design similar to the approach used in this study uses Bayesian optimization and heuristic querying to elicit car form preference [14]. We build on these prior approaches to create an interactive optimization process based on pairwise decisions and feedback through design modification, which is then used to investigate individual and aggregate-level perceptions of subjective semantic attributes.

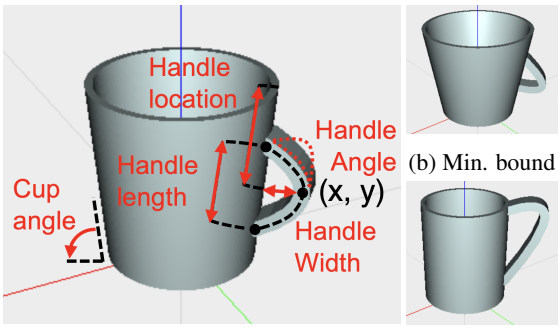
3. METHODS

Pairwise queries, generated actively, were used to find designs that optimize a subjective attribute. These outcomes were evaluated and compared across individual and aggregate models. The interactive optimization method and the study procedure are outlined in the following section.

3.1 Interactive Optimization

3.1.1 Design Example The chosen design example was a drinking mug, while the subjective attribute of interest was how comfortable the mug was to hold. The subjective dimension of "comfortable-to-hold" (vs. a more abstract consideration like "elegance") was considered because there are known aspects of the design, related to variations in how the mug can be held, that are likely to be associated with perceptions and can be understood more clearly. This attribute is not strictly based on visual perception, but since a mug is an everyday object that most participants have picked up and used, it was expected that decisions based on visual information were sufficient. The 3D model of the mug (shown in Figure 1) consisted of a cup with a fixed height, thickness, and bottom radius and a fixed thickness handle created from a Bezier curve with three control points. The mug had five variable parameters: the taper of the cup (cup angle), the distance between the first and last handle control points along the cup surface (handle length), the location of the center of first and last control points along the cup surface (handle location), and the x (handle width) and y (handle angle) positions of the middle control point. These parameters were selected to directly map to how the mug can be held using the handle and the outside of the cup. The parameter bounds were set to extremes that were perceptually reasonable for mugs that exist in reality (shown in Figure 1b and 1c) and these bounds were used to range normalize the design space to a five-dimensional unit hypercube. The variables were treated as continuous within the hypercube, but a small discretization was used during query selection.

3.1.2 Gaussian Process Model A GP model is a surrogate model, specifically a multivariate Gaussian, specified by a mean function and a covariance kernel. The GP used to model the pairwise queries in our study was specified by Chu and Ghahramani, and has been commonly applied to preference learning tasks [43]. Using a probit likelihood, binary observations can be used to infer a latent function (in our case, the participant's perception of the subjective attribute related to the design). Based on Bayes' theorem, the posterior probability func-



(a) Parameter types and labels

(b) Min. bound

(c) Max. bound

FIGURE 1: Design space of parameterized mug with corresponding feature labels presented to participants.

tion, which is the probability of a function given the data, is

$$P(f|D) = \frac{P(f)}{P(D)} P(D|f) \tag{1}$$

where $P(f)$ is the prior (probability of the function), $P(D)$ is the marginal (probability of the data being observed), and $P(D|f)$, the likelihood, is the joint probability of observing the preference data given their latent function values,

$$P(D|f) = \prod_{k=1}^{m} P(v_k \succ u_k | f(v_k), f(u_k)). \tag{2}$$

The probability in the likelihood above is 1 if $f(v_k) \geq f(u_k)$ and 0 otherwise, in the ideal case, but a more tolerant formulation assumes that the latent functions are contaminated with noise that follows a Gaussian distribution. Therefore, at each pairwise decision, the model can maintain an estimate of the participants' utility function, with uncertainty, over a set of points. The maximum posterior mean, the point that maximizes the mean of the estimated functions, can be used to approximate the "best" point throughout the optimization process.

An implementation from the BoTorch Python library was used to fit the model and sample from its posterior at each step[1]. The BoTorch implementation uses a Laplace approximation of the posterior and a radial basis function kernel (also known as the squared exponential kernel) as the covariance function [44].

3.1.3 Active Query Generation There are several options for actively determining the next query to present to users in order to efficiently model their preference decisions. Our approach was adapted from the algorithm used by Tucker et al. for exoskeleton gaits [15], which is based on Thompson sampling and one-dimensional subspaces. Similar line-search approaches have been utilized in other domains such as visual design [40].

[1] PairwiseGP from https://botorch.org/

While other common acquisition functions (e.g., expected improvement or upper confidence bound) were considered, this approach was chosen due to its tractability for balancing exploration and exploitation at higher dimensions and variable levels. The hyperparameter (m) for discretizing the design space was set to be as small as possible while maintaining an evaluation time reasonable for human-computer interaction. A brief summary is included below:

1. The initial comparison is collected using random points (D_i).
2. The model is updated with the initial comparison (D_i). If there is no recorded "best" point already, the one selected initially is set as the best point (B_i).
3. A function (f_s) is sampled from the posterior.
4. A line (L) in a random direction through the best point B_i in the design-space hypercube $[0,1]^d$ is found ($d = 5$ here).
5. The function f_s is evaluated over this line L and observed data D, discretized by a hyperparameter m ($m = 0.005$ here).
6. The point maximizing f_s is presented as the next query (N_i).
7. The new best point B_{i+1} is the one that maximizes the posterior mean (f_μ) over line L and observed data D.
8. The model is updated with the new comparison or from the comparison/s generated from feedback (D_{i+1}).
9. The process is repeated from 3 with the new best point B_{i+1}.
10. After all trials, the final "best" point is set as the "comfort-optimized" design (B_{final}).

Co-active feedback was included as an alternative to direct pairwise selection to improve data quality by mitigating cases when people are unable to perceive small visual differences. Feedback was incorporated in our study through eliciting higher-level design modifications (straighter or more angled cup, taller or shorter handle, wider or narrower handle, move handle up or down, make handle bigger or smaller), though these descriptions may be difficult to specify in more complex cases. The modification enacted by these feedback options was a 10% increase or decrease in the single parameter value (or two parameter values in the case of two feedback options: increasing or decrease "handle size"). Since feedback was given with reference to both designs, it was incorporated as a preference over both of the designs. In our specific implementation, if the feedback was out of bounds, the preference was recorded as the reverse with reference to the side that was selected. Bounds were included in this study for simplicity of normalizing inputs for the model. The feedback mechanism could remove the need for the design space to be bounded strictly if alternative approaches are found for normalization. There are a couple limitations of the active querying method implemented. First, it requires the variables to be continuous, which is not always the case for complex design spaces. Second, there is the possibility for repeat comparisons if the model does not find a better query point along the randomly selected line, which was not accounted for in our study.

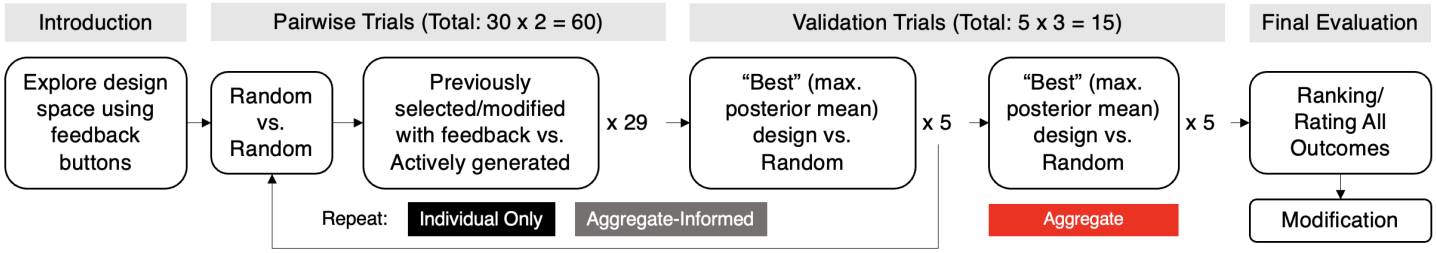


FIGURE 2: Study procedure followed by each participant. The pairwise trials were repeated twice, once with a model containing only individual data and once with a model containing both aggregate and individual data. Three outcomes were compared: the result of the adaptive model and the result of a non-adaptive model with only aggregate data.

3.2 Study Procedure

3.2.1 Participants Data from 56 participants (29 women, 24 men, 3 nonbinary) were collected with approval from an Institutional Review Board. These participants were recruited from university mailing lists that primarily consist of engineering undergraduate or graduate students and as such, do not necessarily represent a general population.

3.2.2 Pairwise Trials and Evaluation Participants made decisions over two sets of 30 pairwise trials. The number of trials was determined based on prior work, though it is possible fewer queries could be used. The difference between the two sets was solely the data used to initialize the model. In one condition, the model had no initial data and was updated using the decisions that the participant provided in the moment (referred to as Individual Only or I). In the other condition, the model was initialized with comparisons generated by a simulation or the first group of participants, and then updated using the participant's in-the-moment decisions (called Aggregate-Informed or A). No time discounting or weighting of more recent answers was included. The condition order was randomly determined when the participant started the study. After each set of trials, during which the model was updated and queries were actively presented, participants were asked to choose between the comfort-optimized design (B_{final}) and a random design for a set of 5 validation trials. They were not explicitly aware of the transition between model updating and validation. At the end, participants also completed 5 validation trials for a third condition (called Aggregate or G) where they selected between the maximum posterior mean design from the Aggregate model (evaluated over a line L and observed data D) and a random design. The Aggregate model was initialized with comparisons generated by either simulation or other prior participants, but not updated with any participant-specific data. All three conditions are summarized in Table 1.

After all trials were completed (75 in total), participants compared the comfort-optimized designs (B_{final} for each adaptive models and the maximum posterior mean for the non-adaptive model) from each condition and provided a rating and ranking. Then, they were allowed to indicate any changes they would make to those designs. Finally, they were directed to a survey where they answered several questions about their decision making. Figure 2 shows the outline of the study.

TABLE 1: Data used to initialize and update the models in each condition (I, A, G) and for each group (1, 2). I and A result in varying comfort-optimized designs across participants and groups while G only results in varying comfort-optimized designs across groups.

Model	Input Data	Updating Data
Individual Only (I)	(1) None (2) None	Pairwise trials
Aggregate-Informed (A)	(1) Simulation (2) Group 1 Individual	Pairwise trials
Aggregate (G)	(1) Simulation (2) Group 1 Individual	None

The study was conducted in two stages. In the first stage, the Aggregate-Informed model was initialized with comparisons which were generated by a simulation. This simulation generated comparisons using two utility functions that would be optimized by minimizing the distance to two design points (perceptually different, but "reasonable" looking designs). One of the two utility functions was used randomly with a random amount of noisiness to represent each "participant." Data from 25 participants was collected in this stage (Group 1). In the second stage, the data from the Individual Only condition of Group 1 was used for the Aggregate and Aggregate-Informed models of Group 2. Data from 31 participants was collected in this stage (Group 2). Although a group of 25 may not be sufficient to represent a true "crowd", increasing the amount of data in the Aggregate-Informed adaptive model increases the computation time per trial and therefore, investigation of crowd size is left to

future consideration.

3.2.3 Interface The custom web interface (developed using Flask and hosted on a Google Compute Engine virtual machine with 8vCPU and 16GB memory) for the pairwise trials is shown in Figure 3. Inspired by the interface in [21], participants could see instructions, followed by side-by-side 3D models of the two designs being compared. Each 3D model was dynamically rendered (using OpenJSCAD and three.js) during the data collection, similar to [9]. There were two buttons to select either option and a third button to provide a design modification. The third button revealed 12 higher-level options for this feedback (corresponding to increases or decreases in parameter values), shown in Figure 3. Before the pairwise trials, participants had the opportunity to explore the design space to better understand the meaning behind these feedback options (see Figure 2). After the pairwise trials, participants conducted a rating/ranking and indicated any modifications they would make to the optimal designs, after which they were directed to a survey.

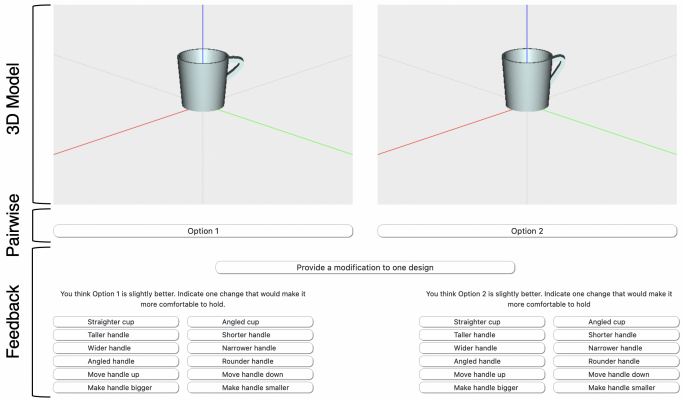


FIGURE 3: Participants made pairwise choices using the interface above. If neither design was perceived as comfortable to hold, a third button revealed choices to suggest a modification.

4. RESULTS

Several models, resulting in comfort-optimized designs, were produced through a human subjects study (N = 25 and N = 31). These models were provided different initial information, either none, simulated aggregate data, or real aggregate data, to guide the interactive optimization. Outcomes from the interactive optimization and a non-adaptive aggregate model were compared to understand the impact of individualizing the computational representations of subjective attributes.

4.1 Evaluating interactive optimization outcomes

The outcomes (each participant's unique comfort-optimized design from the two adaptive models and the group comfort-optimized design for the one non-adaptive model) were evaluated for whether they successfully aligned with participants' perceptions. Both a hit rate obtained separately for each condition and a rating that directly compared all conditions were used for evaluation.

4.1.1 Hit rate The hit rate refers to how often a participant selected the model-predicted comfort-optimized design vs. a random design for the five validation comparisons. Therefore, the hit rate can help demonstrate whether a model can achieve a comfort-optimized outcome relative to the rest of the considered design space. Mann-Whitney U tests show a significant difference in hit rate across groups for the Aggregate ($U = 219.5, p = 0.003$, 1: $Mdn = 0.8$, 2: $Mdn = 1$) and Aggregate-Informed ($U = 122.0, p = 9.25 \times 10^{-7}$, 1: $Mdn = 0.6$, 2: $Mdn = 1$) conditions but not the Individual Only condition ($U = 430.5, p = 0.39$, 1: $Mdn = 1$, 2: $Mdn = 1$). *The results indicate that using simulated data to initialize the model negatively impacts outcomes, with participants often able to find a random design better aligned to their perceptions.*

A non-parametric Friedman test (similar to a repeated measures ANOVA) shows a significant difference between the hit rates across the conditions for Group 2 ($\chi^2(2, N = 31) = 6.2, p = 0.044$). Post-hoc Wilcoxon signed-rank tests with a Bonferroni correction ($\alpha_{new} = 0.05/3 = 0.017$) reveal that this may be driven primarily by a difference between the Individual Only and Aggregate-Informed condition ($W = 11.0, p = 0.018$). However, the median hit rates for Group 2 (I: $Mdn = 1, Range = [0.4, 1]$ A: $Mdn = 1, Range = [0.6, 1]$ G: $Mdn = 1, Range = [0.2, 1]$) demonstrate the success of finding comfort-optimized outcomes using the models generally.

4.1.2 Ratings Ratings of design outcomes in each condition are shown in Figure 4 for Group 2, based on participants' answers to how well aligned each design was to their perception of a mug that is comfortable to hold. It should be noted that the ratings were completed by comparing each different condition directly and therefore also constitute a ranking. *Ratings are highest for the comfort-optimized design from the Aggregate-Informed condition* (I: $Mdn = 5$ A: $Mdn = 6$ G: $Mdn = 5$). There are differences between Individual Only and Aggregate-Informed outcome ratings ($Mdn = -1, Range = [-5, 2]$) and Individual Only and Aggregate ratings ($Mdn = -1, Range = [-5, 3]$), where Aggregate-Informed and Aggregate tend to be rated higher Individual Only. The difference between Aggregate-Informed and Aggregate ratings ($Mdn = 1, Range = [-2, 3]$) shows that Aggregate-Informed also tends to be rated higher than Aggregate. Group 2 appears to have some commonality with Group 1, the source of the aggregate data, regarding comfort perceptions. *This commonality, combined with the ability to guide the outcome with individual decisions might explain the high rat-*

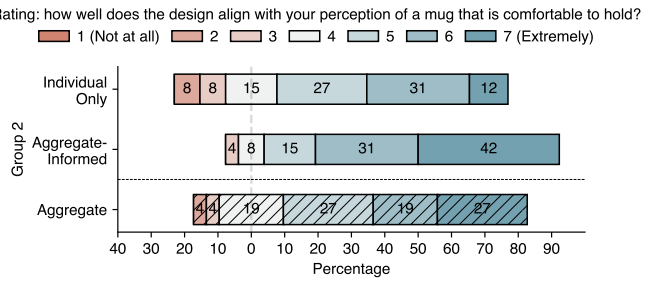


FIGURE 4: Ratings for the Individual Only ($Mdn = 5$), Aggregate-Informed ($Mdn = 6$), and Aggregate outcomes ($Mdn = 5$) in Group 2.

ings and hit rate for the Aggregate-Informed outcomes for Group 2 compared to the other conditions. At the same time, individual participants' rankings (obtained from their comparative ratings) show that *25% (8 of 31) of the participants rank the result of the Individual Only condition as the most aligned with their perception of comfortable. While not the majority, this subset could be considered for application of personalization.*

4.2 Differences reflected by individual vs. aggregate models

The optimization process and outcomes were further analyzed to understand individual and aggregate perceptions of the subjective attribute. Only the data from Group 2 (non-adaptive aggregate model constructed with real data from Group 1) was considered in this analysis.

4.2.1 Exploration For each participant, the designs seen during the interactive trials (Individual Only and Aggregate-Informed) vary because of the adaptive querying. The generalized variance (determinant of the covariance matrix) of all of the designs that were visited throughout the process can give insight into the extent the design space was explored during the process. The generalized variance for designs visited by participants in the Individual Only condition ($Mdn = 1.07 \times 10^{-8}$) is greater ($W = 84.0, p = 0.0013$) than that of the Aggregate-Informed condition ($Mdn = 5.34 \times 10^{-11}$), using a Wilcoxon signed-rank test. The generalized variance of participants' final outcomes is similarly greater for the Individual Only condition (9.72×10^{-7}) than for the Aggregate-Informed condition (1.33×10^{-9}). *Since the Aggregate-Informed condition allows participants to start from a similar "group-level" design, it follows that the spread of designs visited and the corresponding diversity of outcomes for this condition is lower. However, based on the hit rates and ratings, less exploration does not have a negative impact if the starting point is more aligned with human perceptions.*

4.2.2 Outcomes The final outcomes that result from the two interactive conditions (I and A), compared to the Aggregate outcome (G), which involves *no intervention of the individual participant*, are shown in Figure 5 as a 2D projection of the 5-dimensional design space (only for visualization). The examples demonstrate the diversity of participants' final outcomes. Notably, *in many cases the non-adaptive aggregate model may be sufficient*, based on the clustering of number-one ranked individualized outcomes around the group-level outcome. However, *in some cases, participants consider outcomes that are not close to the aggregate as best aligned with their perceptions of comfort.*

As participants were not informed that they were experiencing different conditions, with all trials presented in the exact same way, it is expected that differences are influenced by the initialization data provided to the model or the difference of an individual's preference (e.g., an "extreme" vs. not) from the Aggregate outcome. The standardized Euclidean distance between each individual outcome and the Aggregate outcome demonstrates a measure of this difference for both the Individual Only ($M = 0.53, SD = 0.29, Range = [0.04, 1.20]$) and Aggregate-Informed ($M = 0.31, SD = 0.24, Range = [0.01, 0.88]$) conditions. As expected based on the study design and shown by the generalized variance over the comfort-optimized outcomes, the average distance from the Aggregate is higher for the Individual Only condition than for the Aggregate-Informed condition. There is some evidence of a negative correlation between this Euclidean distance and the Aggregate hit rate ($\rho_s = -0.43, p = 0.02$). However, there is no evidence of such a relationship with the Aggregate rating ($\rho_s = -0.19, p = 0.31$). *Increases in the distance between an individualized outcome and the aggregate-level outcome appear to be associated with a decreasing hit rate. The hit rate is a relative measure of the outcome, indicating that participants with greater differences are more likely to select a random design over the group-level comfort-optimized design. However, this relationship is not reflected through the rating, an "absolute" outcome measure. This may be explained by participants being satisfied enough with several options within the considered design space.*

4.3 Parameter-level differences reveal driving factors of individual differences

Though the designs were considered holistically, the outcomes can also be examined to investigate where variations were most prevalent and whether certain features drove individual differences. Figure 6 shows that certain parameter values were common across the final outcomes for all participants, such as a longer handle. Other parameters, such as the handle angle demonstrate more variety. The parameter with the highest standard deviation among best designs is handle location ($SD = 0.36$) for the Individual Only condition and handle width ($SD = 0.28$) for Aggregate-Informed condition. The parameter that differs the

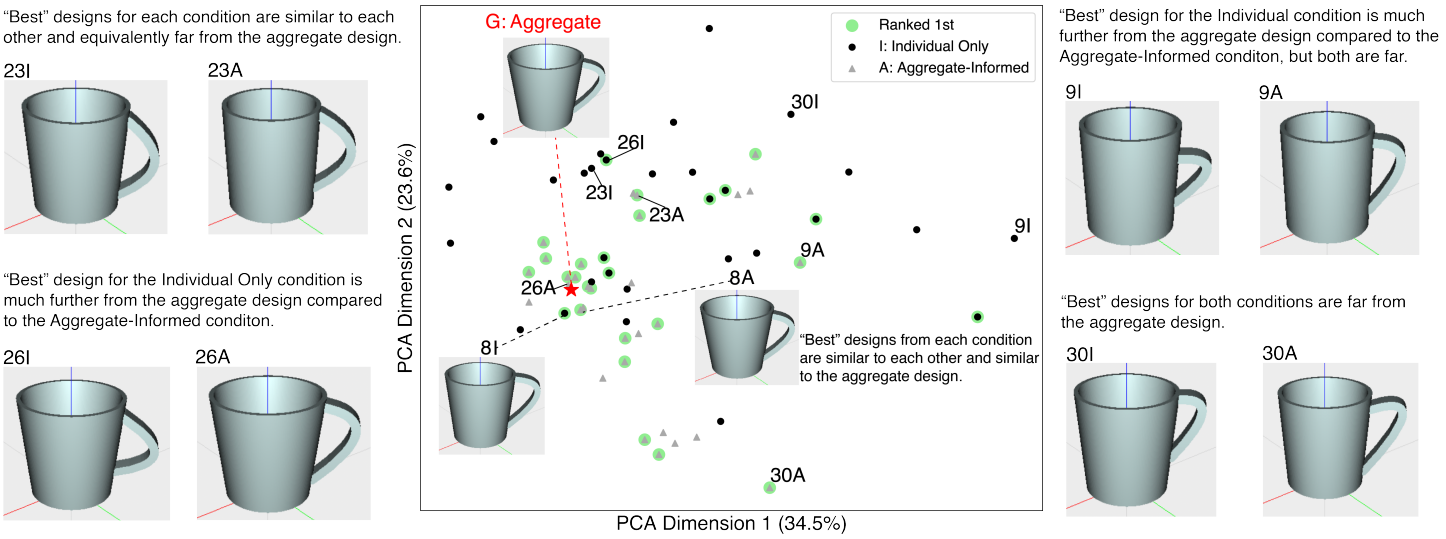


FIGURE 5: Visualization and examples of comfort-optimized ("best") design outcomes from both sets of pairwise trials vs. the aggregate (Group 2 only). The (green) highlights mark if that design was ranked by the participant as the design that most aligned with their perceptions of comfort.

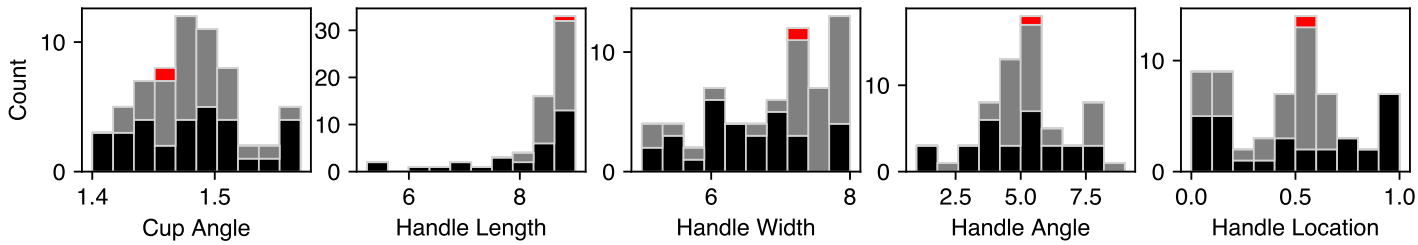


FIGURE 6: Parameters of comfort-optimized designs as determined by optimization in each condition for each participant (Black = Individual Only, Gray = Aggregate-Informed) and the comfort-optimized Aggregate design from Group 1 (Red).

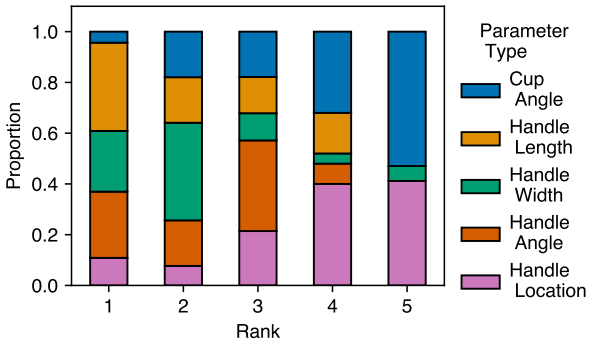


FIGURE 7: Self-reported rankings on the importance of each parameter for participants' decision making (1 = Most important).

least is handle length ($SD = 0.21$ and $SD = 0.05$) in both cases.

Self-reported rankings of how important each parameter was to the participant when making their decisions (ties allowed) are shown in Figure 7. Participants demonstrate variety in their importance rankings. The parameter that is considered the most important by a plurality of participants is handle length. Correspondingly, the smallest standard deviation in outcomes for both conditions is the handle length, which demonstrates one common feature across the group. The handle width is considered the second most important by a plurality of participants. This feature also exhibits the highest standard deviation amongst participants' best designs from the Aggregate-Informed condition. Thus, in this example, it appears that varying the handle width may have driven many of the individual differences.

5. DISCUSSION

To allow computational methods to better align with human perception, it is important to understand how human perception can be best embedded into optimization processes. In this work, we explore how interactive optimization methods lead to outcomes that are well or not well-aligned with humans' percep-

tions of the subjective attribute of comfort. Furthermore, the use of a GP allows for the estimation of a surrogate function which can represent such a semantic scale in relation to design features, though this function is more representative of positive values (more comfortable) than negative ones (less comfortable). In addition, we identify how different outcomes might arise when aggregating this perception across a group of individuals. We find commonalities in human perception within a group that can be leveraged, and find that addressing subtle differences of perceptual preference can be beneficial for satisfaction in outcomes.

5.1 Interactive optimization can capture human perceptions of a subjective attribute

The results of this study indicate that the interactive models are able to produce outcomes that capture perceptions of the subjective attribute considered here reasonably well, based on the high hit rates across both Group 1 and 2. Alignment of individual outcomes with perceptions of comfort is rated relatively highly, with medians of 6 and 5 out of 7. Notably, when the data is simulated and does not match how humans make real decisions (e.g., prioritizing a specific design attribute like handle length), including aggregate data is detrimental, as participants are unable to reach an area of the design space that reflects their perceptions. However, when the data included reflects real human decisions, including this information helps improve upon the individual models, which already perform relatively well. Prior work implies through simulated experiments that changing the initial guess based on similar users is only valuable when the optimally-preferred designs are clustered [26]. Our empirical results support this based on the difference in results from initialization with simulated vs. real data. In general, the example considered here as well as bias in the participant pool likely induces more subtle individual differences in perception, whereas different examples may elicit larger differences that may lessen the benefits of including aggregate data.

5.2 Individualizing models of the perceptual attribute can improve satisfaction with optimized outcome

The use of individual-level models to enable adaptive design exploration along subjective semantic attributes is examined. Design-relevant computational approaches such as semantic shape editing rely on the creation of large aggregate mappings of semantics to geometries [18]. Here, we find that such an approach can work well, but it is possible to move towards personalizing the semantic mappings in order to capture subtle differences across individuals' perceptions. Prior work within the design field has successfully moved towards both crowd-based and adaptive, personalized methods in the context of inspirational stimuli [45, 46]. In this study, the individual models are able to lead to satisfactory outcomes with relatively few queries. However, results show that incorporating real group-level data can lead to better alignment with participants' perceptions, as shown by both their decisions and self-reported measures. Therefore, in practice, individualized models may be more useful when there are highly diverging views of the dimension (e.g., particularly abstract concepts) than in the case considered here.

5.3 Guidance during interactive optimization

While more of the design space is explored during the Individual Only condition, it appears that benefit of the Aggregate-Informed condition is its ability to reach a general consensus of a comfortable mug and then allow adaptation to more subtle user preferences. This adaptation is likely enabled by guidance towards parameters that are more important from prior data (which is unavailable when the model starts from scratch), supported by the parameter-level analysis. Chan et al. find in a study comparing human and optimizer-led design, that though performance-related outcomes can be improved, people lose agency and ownership when they are being guided by an optimizer [47]. Some of this can be mitigated by allowing people to provide co-active feedback, like in the form of design modifications here. People may have been more satisfied with outcomes because they had the option for active guidance rather than only passive evaluation. People may also feel frustrated if they do not understand Bayesian optimization, which trades off exploration and exploitation [47]. Thus, it may appear the optimizer is giving worse examples when it is simply trying to gain more information. Open-ended comments from the survey in our study indicate that some participants felt like they could see the impact of their decisions and feedback throughout the process, while others felt frustrated if they felt like their decisions did not make a difference or if the differences were not visually perceivable.

5.4 Limitations and Future Work

Some limitations and further work should be considered. Because the adaptive querying uses a random line search and the optimization uses sampling, there is a degree of randomness in what participants saw during the study. Furthermore, the complexity of the design may impact the optimization performance and the method has not been tested beyond six dimensions [15]. Some limitations also relate to the design representation. Most notably, it is known that evaluations can differ between digital and physical models [48]. However, the approach can be used to narrow the design space down for more expensive design representations to be evaluated. Additionally, in this study, the design was parameterized into five features that were assumed to be relatively important for the subjective attribute being considered. The surrogate function for each participant could likely be used to optimize a mug of a different shape for "comfort" as long as it can be parameterized by the features represented here (e.g. handle length, handle width). However, this is not inclusive to many other features people may have considered, which may account for further inter-individual differences. In our case, the adaptive querying was conducted using design features that mapped

directly to low-level parameters. Participants were allowed to provide slightly abstract feedback, but an interesting area of future investigation might be to consider cases where higher-level conceptual feedback does not have a one-to-one or one-to-two mapping with the low-level variable. It should also be noted that the approach taken here may have to be modified if the adjectives describing the semantic pair are very different (e.g., traditional to elegant) compared to a quantity (less or more comfortable).

Prior work notes that human steering can impact optimization if information about how the process works is provided [49]. Therefore, it is possible that if participants were able to understand the impact of their actions on the optimization process, they would reach better outcomes. Furthermore, with the increasing prevalence of generative models, efforts have been taken to guide them to better align with subjective evaluations in the context of design [50–52]. The approach taken in the study conducted here could be used in conjunction with generative models to generate outputs that are aligned with a specific, personalized semantic attribute. Such future advances can eventually lead to tools that enable designers to explore a vast design space while not having to sacrifice their individualized styles.

6. CONCLUSION

In this paper, interactive Bayesian optimization is utilized to capture and investigate human perceptions of a subjective attribute. We provide insight into the ability to capture both an aggregate-level representation and show how subtle individual differences may result in different outcomes and satisfaction with these outcomes. The results show that an aggregate-level model can represent human perception of preferences well, but that including individual differences results in even better alignment. While people may share some commonalities in their perceptions, in order to interact with designs at a semantic level, it may be useful to enable individualized semantic mappings of subjective attributes to design features. The Gaussian Process approach, which results in a design optimized for the subjective attribute for each person, estimates a surrogate function connecting features to perceptions, which can then be used in further applications.

ACKNOWLEDGMENT

This work has been supported by the Regents of University of California and NSF (2145432-CAREER). The findings represent the views of the authors and not necessarily those of the sponsors.

REFERENCES

[1] Krippendorff, Klaus and Butter, Reinhart. "Product semantics-exploring the symbolic qualities of form." *Innovation* Vol. 3 (1984): pp. 4–9.

[2] Petiot, Jean-François and Yannou, Bernard. "Measuring consumer perceptions for a better comprehension, specification and assessment of product semantics." *International Journal of Industrial Ergonomics* Vol. 33 No. 6 (2004): pp. 507–525.

[3] Orsborn, Seth, Cagan, Jonathan and Boatwright, Peter. "Quantifying aesthetic form preference in a utility function." *J. Mech. Des.* Vol. 131 No. 6 (2009): p. 061001.

[4] Kelly, Jarod C, Maheut, Pierre, Petiot, Jean-François and Papalambros, Panos Y. "Incorporating user shape preference in engineering design optimisation." *Journal of Engineering Design* Vol. 22 No. 9 (2011): pp. 627–650.

[5] Ren, Yi, Burnap, Alex and Papalambros, Panos. "Quantification of perceptual design attributes using a crowd." *DS 75-6: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol. 6: Design Information and Knowledge, Seoul, Korea, 19-22.08. 2013*. 2013.

[6] Sylcott, Brian and Cagan, Jonathan. "Modeling aggregate choice for form and function through metaconjoint analysis." *J. Mech. Des.* Vol. 136 No. 12 (2014): p. 124501.

[7] Goucher-Lambert, Kosa and Cagan, Jonathan. "The impact of sustainability on consumer preference judgments of product attributes." *J. Mech. Des.* Vol. 137 No. 8 (2015): p. 081401.

[8] Valencia-Romero, Ambrosio and Lugo, José E. "Part-worth utilities of Gestalt principles for product esthetics: a case study of a bottle silhouette." *J. Mech. Des.* Vol. 138 No. 8 (2016): p. 081102.

[9] Kang, Namwoo, Ren, Yi, Feinberg, Fred and Papalambros, Panos. "Form+ function: Optimizing aesthetic product design via adaptive, geometrized preference elicitation." *arXiv preprint arXiv:1912.05047* (2019).

[10] Lin, David Chuan-En and Martelaro, Nikolas. "Learning Personal Style from Few Examples." *Designing Interactive Systems Conference 2021*: p. 1566–1578. 2021. Association for Computing Machinery, New York, NY, USA. DOI 10.1145/3461778.3462115.

[11] Poirson, E and Petiot, J-F. "Interactive genetic algorithm to collect user perceptions. Application to the design of stemmed glasses." *Nature-Inspired Methods for Metaheuristics Optimization: Algorithms and Applications in Science and Engineering* (2020): pp. 35–51.

[12] Bıyık, Erdem, Huynh, Nicolas, Kochenderfer, Mykel J and Sadigh, Dorsa. "Active preference-based gaussian process regression for reward learning." *Robotics: Science and Systems 2020*. 2020.

[13] Williams, Christopher KI and Rasmussen, Carl Edward. *Gaussian processes for machine learning*. Vol. 2. MIT Press, Cambridge, MA (2006).

[14] Ren, Yi and Papalambros, Panos Y. "A design preference elicitation query as an optimization process." *J. Mech. Des.*

Vol. 133 No. 11 (2011).

[15] Tucker, Maegan, Cheng, Myra, Novoseller, Ellen, Cheng, Richard, Yue, Yisong, Burdick, Joel W and Ames, Aaron D. "Human preference-based learning for high-dimensional optimization of exoskeleton walking gaits." *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*: pp. 3423–3430. 2020. IEEE.

[16] Nielsen, Jens Brehm Bagger, Nielsen, Jakob and Larsen, Jan. "Perception-based personalization of hearing aids using Gaussian processes and active learning." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* Vol. 23 No. 1 (2014): pp. 162–173.

[17] Lun, Zhaoliang, Kalogerakis, Evangelos, Wang, Rui and Sheffer, Alla. "Functionality Preserving Shape Style Transfer." *ACM Trans. Graph.* Vol. 35 No. 6 (2016). DOI 10.1145/2980179.2980237.

[18] Yumer, Mehmet Ersin, Chaudhuri, Siddhartha, Hodgins, Jessica K and Kara, Levent Burak. "Semantic shape editing using deformation handles." *ACM Transactions on Graphics (TOG)* Vol. 34 No. 4 (2015): pp. 1–12.

[19] Nagamachi, Mitsuo. "Kansei engineering: a new ergonomic consumer-oriented technology for product development." *International Journal of industrial ergonomics* Vol. 15 No. 1 (1995): pp. 3–11.

[20] Reid, Tahira N, Gonzalez, Richard D and Papalambros, Panos Y. "Quantification of perceived environmental friendliness for vehicle silhouette design." *J. Mech. Des.* Vol. 132 No. 10 (2010): p. 101010.

[21] Zintgraf, Luisa M., Roijers, Diederik M., Linders, Sjoerd, Jonker, Catholijn M. and Nowé, Ann. "Ordered Preference Elicitation Strategies for Supporting Multi-Objective Decision Making." *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*: p. 1477–1485. 2018. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.

[22] Petiot, Jean-François and Grognet, Stephane. "Product design: a vectors field-based approach for preference modelling." *Journal of Engineering Design* Vol. 17 No. 03 (2006): pp. 217–233.

[23] Petiot, Jean-François and Dagher, Antoine. "Preference-oriented form design: application to cars' headlights." *International Journal on Interactive Design and Manufacturing (IJIDeM)* Vol. 5 (2011): pp. 17–27.

[24] Nandy, Ananya and Goucher-Lambert, Kosa. "Do human and computational evaluations of similarity align? An empirical study of product function." *J. Mech. Des.* Vol. 144 No. 4 (2022): p. 041404.

[25] Carroll, J Douglas and Chang, Jih-Jie. "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition." *Psychometrika* Vol. 35 No. 3 (1970): pp. 283–319.

[26] Ren, Yi and Papalambros, Panos Y. "On design preference elicitation with crowd implicit feedback." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 45028: pp. 541–551. 2012. American Society of Mechanical Engineers.

[27] Wan, Jie and Krishnamurty, Sundar. "Learning-based preference modeling in engineering design decision-making." *J. Mech. Des.* Vol. 123 No. 2 (2001): pp. 191–198.

[28] Thurston, Deborah L. "Real and misconceived limitations to decision based design with utility analysis." *J. Mech. Des.* Vol. 123 No. 2 (2001): pp. 176–182.

[29] Valencia-Romero, Ambrosio and Lugo, José E. "An immersive virtual discrete choice experiment for elicitation of product aesthetics using Gestalt principles." *Design Science* Vol. 3 (2017): p. e11.

[30] Reid, Tahira N, Frischknecht, Bart D and Papalambros, Panos Y. "Perceptual attributes in product design: Fuel economy and silhouette-based perceived environmental friendliness tradeoffs in automotive vehicle design." *J. Mech. Des.* Vol. 134 No. 4 (2012): p. 041006.

[31] Sylcott, Brian, Michalek, Jeremy J and Cagan, Jonathan. "Towards understanding the role of interaction effects in visual conjoint analysis." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 55881: p. V03AT03A012. 2013. American Society of Mechanical Engineers.

[32] Burnap, Alexander, Hartley, Jeffrey, Pan, Yanxin, Gonzalez, Richard and Papalambros, Panos Y. "Balancing design freedom and brand recognition in the evolution of automotive brand styling." *Design Science* Vol. 2 (2016): p. e9. DOI 10.1017/dsj.2016.9.

[33] Tseng, Ian, Cagan, Jonathan and Kotovsky, Kenneth. "Concurrent optimization of computationally learned stylistic form and functional goals." *J. Mech. Des.* Vol. 134 No. 11 (2012): p. 111006.

[34] Burnap, Alex, Pan, Yanxin, Liu, Ye, Ren, Yi, Lee, Honglak, Gonzalez, Richard and Papalambros, Panos Y. "Improving design preference prediction accuracy using feature learning." *J. Mech. Des.* Vol. 138 No. 7 (2016): p. 071404.

[35] Lepird, John R, Owen, Michael P and Kochenderfer, Mykel J. "Bayesian preference elicitation for multiobjective engineering design optimization." *Journal of Aerospace Information Systems* Vol. 12 No. 10 (2015): pp. 634–645.

[36] Desmedt, Nicolas, Iliopoulou, Vicky, Lopez, Carlos and De Grave, Kurt. "Active preference learning in product design decisions." *Procedia CIRP* Vol. 100 (2021): pp. 277–282.

[37] Brochu, Eric, Cora, Vlad M and De Freitas, Nando. "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical

reinforcement learning." *arXiv preprint arXiv:1012.2599* (2010).

[38] Tao, Siyu, Van Beek, Anton, Apley, Daniel W and Chen, Wei. "Multi-model Bayesian optimization for simulation-based design." *J. Mech. Des.* Vol. 143 No. 11 (2021).

[39] Iyer, Akshay, Yerramilli, Suraj, Rondinelli, James M, Apley, Daniel W and Chen, Wei. "Descriptor aided Bayesian optimization for many-level qualitative variables with materials design applications." *J. Mech. Des.* Vol. 145 No. 3 (2023): p. 031701.

[40] Koyama, Yuki, Sato, Issei, Sakamoto, Daisuke and Igarashi, Takeo. "Sequential line search for efficient visual design optimization by crowds." *ACM Transactions on Graphics (TOG)* Vol. 36 No. 4 (2017): pp. 1–11.

[41] Koyama, Yuki, Sato, Issei and Goto, Masataka. "Sequential gallery for interactive visual design optimization." *ACM Transactions on Graphics (TOG)* Vol. 39 No. 4 (2020): pp. 88–1.

[42] Koyama, Yuki and Goto, Masataka. "BO as Assistant: Using Bayesian Optimization for Asynchronously Generating Design Suggestions." *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*: pp. 1–14. 2022.

[43] Chu, Wei and Ghahramani, Zoubin. "Preference Learning with Gaussian Processes." *Proceedings of the 22nd International Conference on Machine Learning*: p. 137–144. 2005. Association for Computing Machinery, New York, NY, USA. DOI 10.1145/1102351.1102369.

[44] Balandat, Maximilian, Karrer, Brian, Jiang, Daniel R., Daulton, Samuel, Letham, Benjamin, Wilson, Andrew Gordon and Bakshy, Eytan. "BOTORCH: A Framework for Efficient Monte-Carlo Bayesian Optimization." *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020. Curran Associates Inc., Red Hook, NY, USA.

[45] Goucher-Lambert, Kosa and Cagan, Jonathan. "Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation." *Design Studies* Vol. 61 (2019): pp. 1–29.

[46] Goucher-Lambert, Kosa, Gyory, Joshua T, Kotovsky, Kenneth and Cagan, Jonathan. "Adaptive inspirational design stimuli: using design output to computationally search for stimuli that impact concept generation." *J. Mech. Des.* Vol. 142 No. 9 (2020): p. 091401.

[47] Chan, Liwei, Liao, Yi-Chi, Mo, George B, Dudley, John J, Cheng, Chun-Lien, Kristensson, Per Ola and Oulasvirta, Antti. "Investigating positive and negative qualities of human-in-the-loop optimization for designing interaction techniques." *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*: pp. 1–14. 2022.

[48] Häggman, Anders, Tsai, Geoff, Elsen, Catherine, Honda, Tomonori and Yang, Maria C. "Connections between the design tool, design attributes, and user preferences in early stage design." *J. Mech. Des.* Vol. 137 No. 7 (2015): p. 071408.

[49] Colella, Fabio, Daee, Pedram, Jokinen, Jussi, Oulasvirta, Antti and Kaski, Samuel. "Human strategic steering improves performance of interactive optimization." *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*: pp. 293–297. 2020.

[50] Yuan, Chenxi, Marion, Tucker and Moghaddam, Mohsen. "DDE-GAN: Integrating a Data-Driven Design Evaluator Into Generative Adversarial Networks for Desirable and Diverse Concept Generation." *J. Mech. Des.* Vol. 145 No. 4 (2023). DOI 10.1115/1.4056500. 041407.

[51] Yuan, Chenxi, Marion, Tucker and Moghaddam, Mohsen. "Leveraging End-User Data for Enhanced Design Concept Evaluation: A Multimodal Deep Regression Model." *J. Mech. Des.* Vol. 144 No. 2 (2021). DOI 10.1115/1.4052366. 021403.

[52] Jiang, Zhoumingju, Wen, Hui, Han, Fred, Tang, Yunlong and Xiong, Yi. "Data-driven generative design for mass customization: A case study." *Advanced Engineering Informatics* Vol. 54 (2022): p. 101786. DOI 10.1016/j.aei.2022.101786.