Crowdsourcing for Engineering Design: Objective Evaluations and Subjective Preferences

by

Alexander Burnap

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Design Science)
in The University of Michigan
2016

Doctoral Committee:

Professor Panayiotis Y. Papalambros, Co-Chair
Professor Richard D. Gonzalez, Co-Chair
Assistant Professor Matthew K. Johnson-Roberson
Assistant Professor Honglak Lee

For the objective evaluations that unite our values,

and the subjective preferences that keep them interesting.

ACKNOWLEDGEMENTS

Thank you to all the following people, I am very grateful for the roles you have played in my life, and hopefully I will continue to play a role in yours:

Dato Kighuradze, for mentoring me early on that, "if you understand limits, you understand everything." The grown-up version of that is not limited to math. Will Marotta, the owner of a small mechanic shop that played classic music in engine bays and repaired cars based on need. You were the first to show me how to mix art and technology. Jeff Hartley, the first design scientist I have met at a company. Thank you for years of the long-winded stream-of-consciousness meanderings of the mind and mentoring me without judgment.

Diann Brei, for helping my leadership and the writing algorithm; Colleen, for giving me the unabridged read; Zoran Filipi, Jeff Stein, Dawn Tilbury, and Richard Gerth for helping me figure out what was important to work on with vehicles early on in graduate school. Honglak Lee, for opening up the world of changing representations. Matt Johnson-Roberson, for crowdsourcing but also ethics and work-life balance.

To my DESCI family and ODE family: Yanxin Pan and Ye Liu, you are honestly the best possible team to balance our strengths; Max Yi Ren, your constructive criticism was so important, I am very lucky to have began maturing research-wise with you; Namwoo, you always brought compassion; Emrah you helped keep scope of the world ethics and politics; Bill you brought leg day; Yuqing you brought laughs. Steven, Soodeh, and Anna, good times. Thank you also to Giannis, Carlie, Charlie,

Christina, and Alex Jr. for helping the process.

Panos Papalambros and Rich Gonzalez, my advisors and deepest intellectual mentors: Panos, you have taught me so much, but perhaps "focus" internally and "respect" externally. Thank you for the mentorship that evolved to friendship and academic kinship. Rich, paradoxically, it is the opposite; focus "externally" on other fields as nothing is that new, and "respect" internally. Thank you for taking me in and allowing me to develop, I can truly not have asked for better advisors.

To my meditation friends, Li Nong and Josh Damron, who share respect for Aurelius, Kabat-Zinn, and Batchelder, for continuing to shape my worldview on the human condition and experience the present. To my life friends, thank you from the depth of my heart. I will continue to give you nice wooden cutting boards at your weddings: Nir, John, Joao, Max, Hannah, Lucija, Zheng, David Dai, Oto, Ananda, Hani, Esteban, Clover, Luis, Efren, and Brian.

Lastly, the most important in my life, my family: Mo and Do, you two individually then collectively overcame the most adverse conditions and raised a family with as much integrity and opportunity as possible. I love you very much and hope to one day carry your incredible torch. To my best friend, and brother, Andrew. We have gotten so much closer and I look up to you more than you know. I am not sure I can keep up the 600 mile rule, but I will be a good uncle–dog pound out.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Crowdsourcing for Engineering Design:
Objective Evaluations and Subjective Preferences

by

Alexander Burnap

Co-Chair: Panayiotis Y. Papalambros

Co-Chair: Richard D. Gonzalez

Crowdsourcing enables designers to reach out to large numbers of people who may not have been previously considered when designing a new product, listen to their input by aggregating their preferences and evaluations over potential designs, aiming to improve "good" and catch "bad" design decisions during the early-stage design process. This approach puts human designers–be they industrial designers, engineers, marketers, or executives–at the forefront, with computational crowdsourcing systems on the backend to aggregate subjective preferences (e.g., which next-generation Brand A design best competes stylistically with next-generation Brand B designs?) or objective evaluations (e.g., which military vehicle design has the best situational awareness?). These crowdsourcing aggregation systems are built using probabilistic approaches that account for the irrationality of human behavior (i.e., violations of reflexivity, symmetry, and transitivity), approximated by modern machine learning algorithms and optimization techniques as necessitated by the scale of data (millions of data points, hundreds of thousands of dimensions).

This dissertation presents research findings suggesting the unsuitability of current off-the-shelf crowdsourcing aggregation algorithms for real engineering design tasks due to the sparsity of expertise in the crowd, and methods that mitigate this limitation by incorporating appropriate information for expertise prediction. Next, we introduce and interpret a number of new probabilistic models for crowdsourced design to provide large-scale preference prediction and full design space generation, building on statistical and machine learning techniques such as sampling methods, variational inference, and deep representation learning. Finally, we show how these models and algorithms can advance crowdsourcing systems by abstracting away the underlying appropriate yet unwieldy mathematics, to easier-to-use visual interfaces practical for engineering design companies and governmental agencies engaged in complex engineering systems design.

CHAPTER I

Introduction

1.1 Introduction

Engineering design problems, particularly those addressed by enterprises governed by self-sustaining profit and contribution within a market, involve human designers with the goal of making the best design decisions during the design process for the targeted end user or market segment (Simon, 1969; Cross, 2007; Bayazit, 2004). These human designers make decisions that ultimately affect the final product or service in every step of the design process, from market assessment, to initial conceptualization, to concept selection, to engineering optimization, to manufacturing, and finally to mass distribution and service (Krishnan & Ulrich, 2001).

We take this notion in this dissertation by putting human designers at the forefront– be they industrial designers, engineers, marketers, or executives–with computational *crowdsourcing* systems on the backend that aggregate *objective evaluations* or *subjective preferences* from members in a given crowd, separate from the designer herself, who have relevant information to the design problem. The goal of these computational crowdsourcing systems is to augment the designer's decision-making process by aggregating the crowd's relevant information to further improve "good" design decisions and to catch potentially "bad" design decisions. At the enterprise level, these crowdsourcing systems allow industrial companies and governmental agencies

the opportunity to improve decision making during the early-stage engineering design process, with the promise of significant cost savings relative to late-stage design decisions, both in budget expenditures and time overruns (Oehmen & Seering, 2011).

These crowdsourcing systems are enabled by the scale and reach of the internet, as well as ubiquitous computing and data collection on the crowd member side (e.g., desktop computers and smartphones). Moreover, these systems are built using the language of probability theory and they are approximated by modern data-driven machine learning and optimization algorithms. While we adopt intuition and definition rigor from axiomatic approaches to aggregating the crowd's input, we opt to develop statistical approaches that are more flexible with regards to otherwise irrational human input behavior (e.g., violations of reflexivity, symmetry, and transistivity).

We begin this introduction by laying the groundwork for later discussed research experiments and findings: First, we introduce the notion of "crowdsourcing" as well as our own practical definition for design science. This is backed up by current uses of crowdsourcing in industry, government agencies, and academia; a taxonomy of various types of crowdsourcing processes; and previous practical findings as to crowdsourcing's relative pros and cons. We then detail the framework of design as a decision-making process, the currently dominant paradigm in the way design is conceptualized and subsequently formalized. This framework leads to a very important dichotomy relevant to this work–the difference between *subjective design decisions* and *objective design decisions*. Next, we discuss the potential benefit of crowdsourcing for engineering design, in that it may improve "good" design decisions and catch "bad" design decisions at early stages of the design process, thus enabling the opportunity for design changes that would otherwise be very costly for the enterprise at later stages of the design process.

Second, we move to quantitative models relevant to the formalizations throughout various research contributions in this work. The presentation begins with an

axiomatic formalization of a single evaluator's input during a crowdsourcing process. We discuss issues stemming from limitations imposed by real-world engineering design problems (e.g., incomplete or noisy evaluator input data, and contradictory input data), thus justifying the use of statistical approaches to model a single evaluator. This notion is then extended to the case of multiple evaluators constituting a crowd, in which we detail a number of historically important quantitative models from diverse disciplines including psychometrics, econometrics, and statistics such as test theory, social welfare theory, and discrete choice theory; as well as more recent models from the machine learning and crowdsourcing communities, which often repeat or extend these historic models.

Third, we discuss the research gap between the current quantitative models and the particular demands for their suitability to engineering design. We note that these models do not specifically account for: (1) The relative sparsity of expertise imbued in the crowd for the given design decision in question, a common scenario in even "simple" engineering design; and (2) the "heterogeneity" in the crowd, in which either the types of expertise needed for the design task or the similarity in design preferences for various market segment are not sufficiently modeled. These two shortcomings are major obstacles to practical adoption of quantitative crowdsourcing processes within real engineering design enterprises, particularly as designs are shifting to ever more complexity (e.g., hybrid-electric and autonomous vehicles) as well as customization (e.g., built-to-order vehicles from the manufacturer) in an increasingly diverse and globalized world (e.g., preferences in different geographic markets or market segment demographics).

Fourth, we discuss the general research framework used–a spectrum spanning objective evaluations to subjective preferences–according to the amount of expertise needed for a given design task. Upon this spectrum, we place the five chapters detailing research findings within this dissertation, as well as brief overviews of how

each of these chapters addresses the research gap identified in this first chapter.

1.2 Crowdsourcing in Engineering Design

1.2.1 What is Crowdsourcing?

Crowdsoucing as a concept is not new. As the concept of crowdsourcing deals with the systematic aggregation of input from humans for a shared task, perhaps the earliest formalization of crowdsourcing may indeed be the Athenian democracy in 5th century B.C. Moving forward in history, the earliest formalized notion of crowdsourcing directly relevant to this dissertation is the 1785 treatise on voting theory by the *Marquis de Condorcet* (Condorcet, 1785), in which a paradox showing that there is no formal voting method for more than two alternatives in which a combined vote satisfies a majority of individual votes; in other words, the popular vote is inconsistent; a finding that was further formalized in 1951 by Arrow to later receive the 1972 Nobel Prize in Economics (Arrow, 1951).

Explicit usage of the term 'crowdsoucing' is generally attributed to being first coined in 2006 by a journalist for Wired magazine (Howe, 2006) as a play on the more commonly known term 'outsourcing.' Since then, there have been hundreds of definitions of 'crowdsourcing' in the academic literature, oftentimes tailored to the particular context (Estelles-Arolas & Gonzalez-Ladron-de Guevara, 2012). Given this dissertation's focus on design science, we define crowdsourcing as:

> "The aggregation of input for a given design task from a number of people other than the designer herself, using a systematic aggregation procedure enabled by the reach and scale of the internet and modern computation, with the goal of augmenting the designer's decision-making during the design process."

Thus, while the concept of crowdsourcing is not new, what is new is the *reach* and

scale we now have to interact with evaluators and customers. We give an overview of these enablers in Figure 1.1. In general, the reach and scale of crowdsourcing are due to the internet and modern computational processing. The internet has led to world-wide networking, which has promoted standardized communication protocols as well as creation of online communities relevant for a given design task. This networking is accessible by a number of devices, in particular desktops and smartphones that even in basic forms often have multicore processors and on-board graphics processing. On the client side, these processing architectures allow much higher fidelity 2D and 3D real-time rendering and manipulation of design concepts. On the server side, we have vastly increased computational power via multicore architectures and RAM/VRAM, allowing highly-parallel CPU and GPU computing. The proliferation of these networked devices and subsequent customer use of web and smartphone applications contribute to massive increases in data sizes. The availability of open source machine learning and optimization libraries, which much of the time far surpass the capabilities of proprietary software, enables fast progress across research disciplines to support the development of the new models and algorithms used in this dissertation.

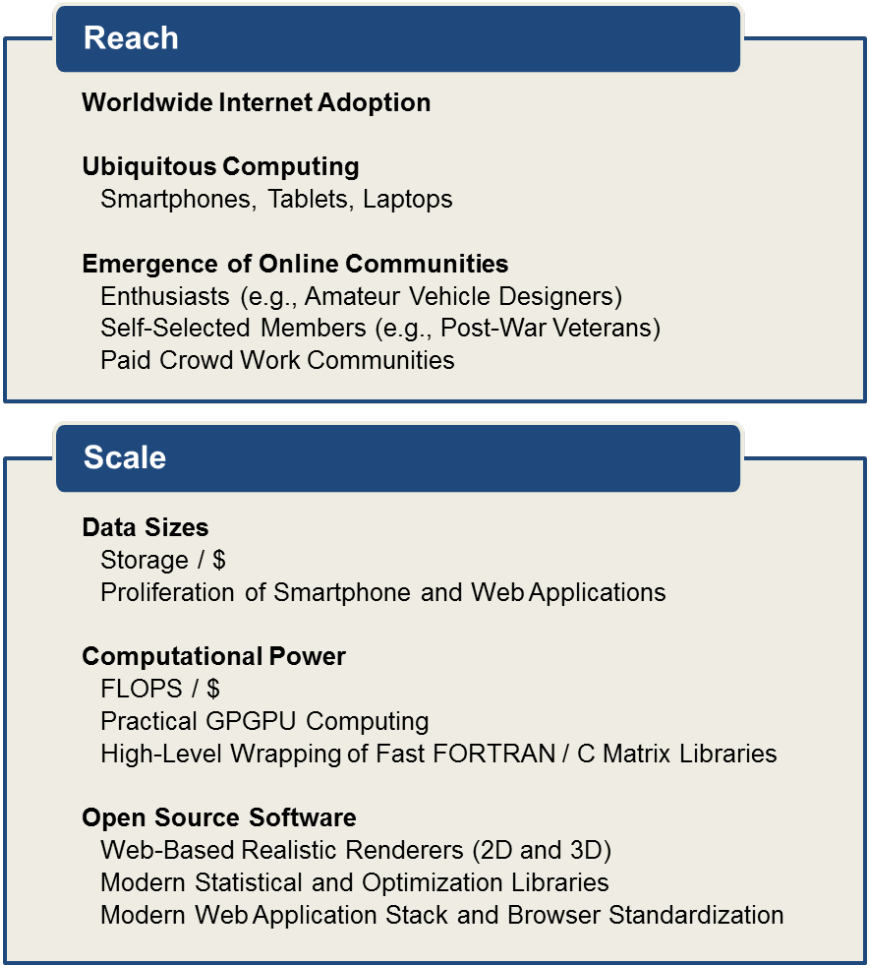


Figure 1.1: Enablers of crowdsourcing as we define in this dissertation; while crowdsourcing as human input aggregation is not new, what is new is the *reach* and *scale* we now have to access evaluators and customers who may have potentially valuable input during the early-stage design process.

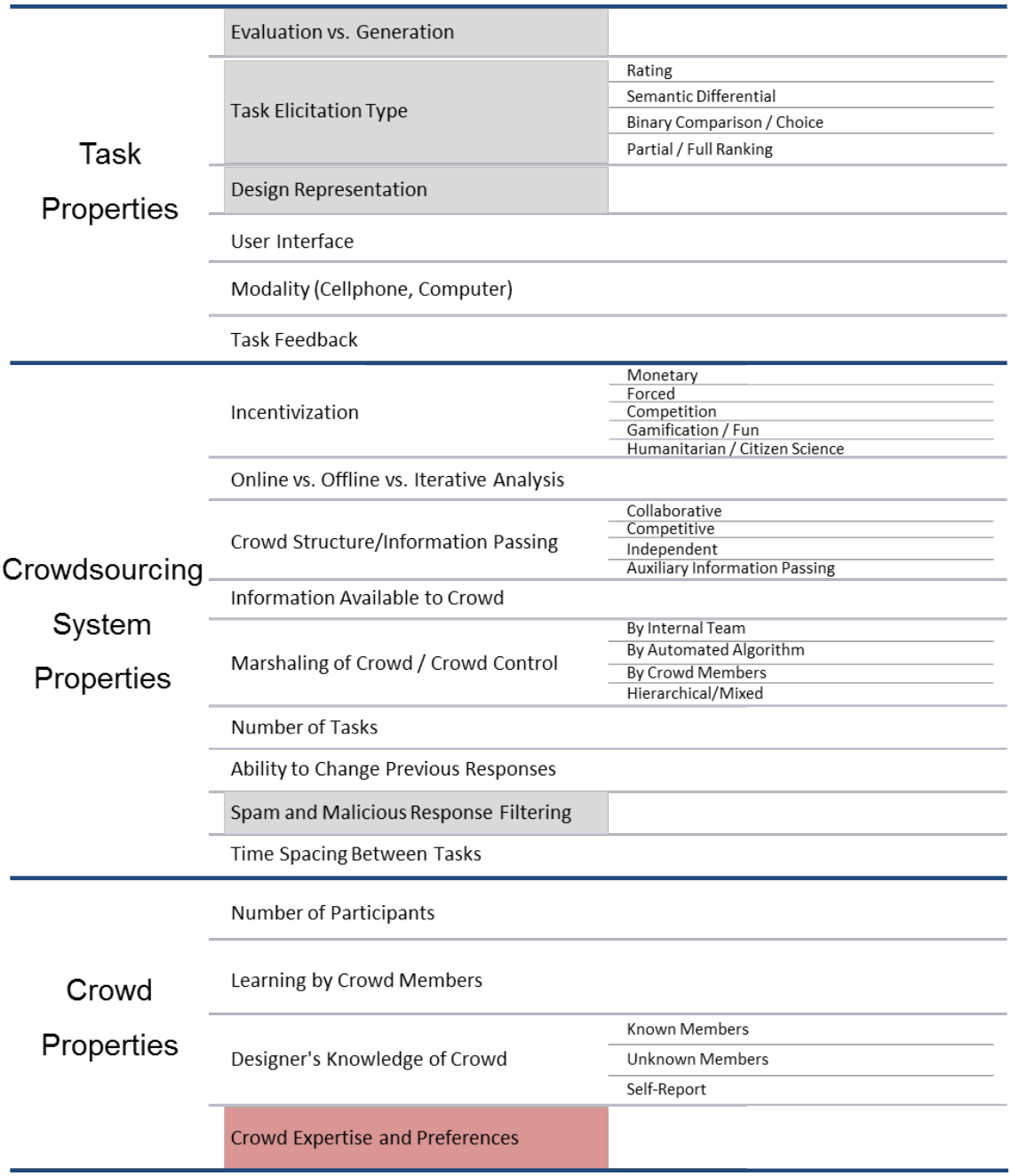


Figure 1.2: Taxonomy of crowdsourcing processes as identified for this dissertation; most identified properties are not considered within this dissertation. Grey shaded boxes show properties that were varied within this dissertation. The red shaded box shows the property that was varied and explicitly studied throughout this dissertation, namely, crowd expertise or crowd preferences.

Taxonomy of Crowdsourcing Processes In our definition we do not restrict who is in the crowd, the number of people in the crowd, or properties of the crowd member themselves. There is, however, a number of properties that differentiate crowdsourcing processes. We give a taxonomy of task properties, crowdsourcing mechanism properties, and crowd properties in Figure 1.2. We focus only on a few properties of a crowdsourcing process, and subsequently, the results of our research contributions are limited to these cases. In short, we only manipulate whether we are using the crowd for design evaluation or generation, the type of task elicitation, the design representation, and the expertise of the crowd.

Numerous major veins of research into aggregation of human input during the engineering design process have preceded the advent of crowdsourcing for engineering design. One such vein is that of coordination of subsystems during systems decomposition, found in the research areas of distributed design or decentralized design (Lee & Whang, 1999; Gurnani & Lewis, 2008), in which a complex design (e.g., vehicle), is decomposed into a number of subsystems (e.g., powertrain, chassis, body), followed by subsequent coordination and recombination. Results show that even in small teams, though slower to converge, group decisions may outperform single expert decisions for certain design tasks (Yang, 2010).

Another major area within the engineering design community relevant to the current work is that of collaborative design with iterative processes such as the Delphi Method, in which a crowd of evaluators, typically restricted to relevant industry or academic professionals, is asked to collectively evaluate a number of possible design options using an iterative survey (Linstone *et al.*, 1975; Dalkey & Helmer, 1963). Such iterations typically involve long time periods while evaluations are collected, and generally involve results of the previous iteration (though not applicable on the first iteration) to spur convergence of the crowd's consensus. While the original Delphi Method is conducted for future technology predictions, work within the engineer-

ing design community has focused on the case of large-scale engineering enterprises, with examples such as risk-assessment of product designs with long development and service lifetimes (Ahn *et al.*, 2014).

Our work is a major departure from many of these previous works, and has the limitation that we do not consider collaboration or other forms of information passing between evaluators. Clearly in a real "crowd," such independence assumptions may be violated, particularly when information passing strongly correlates with expertise, or when collaborative teams have greater expertise than the summation of its individuals (Hong & Page, 2004a). It has been recently shown at enterprises engaged in complex engineering systems design that lack of information passing stifles expertise (McGowan *et al.*, 2013); or that communication is withheld for work negotiation purposes (Austin-Breneman *et al.*, 2014).

1.2.2 The Design Process and Decision-Based Design

The design process refers to the path that a product or service moves through during development, from initial conceptualization to final embodiment and beyond as shown in Figure 1.3. For complex products or services, this path may require thousands of people—from market researchers, designers, engineers, manufacturers, on down—who may be distributed over a number of divisions, companies, and geographical locations. Similarly, for simple products or services, this path may only require one person. Numerous frameworks outlining the design process exist within the academic literature on design, with a number of other design process frameworks having originated within industrial companies or governmental agencies. Surveys of these frameworks show that these frameworks are less a matter of standardization, but more niche-based depending on the complexity and the particular industry (Krishnan & Ulrich, 2001).

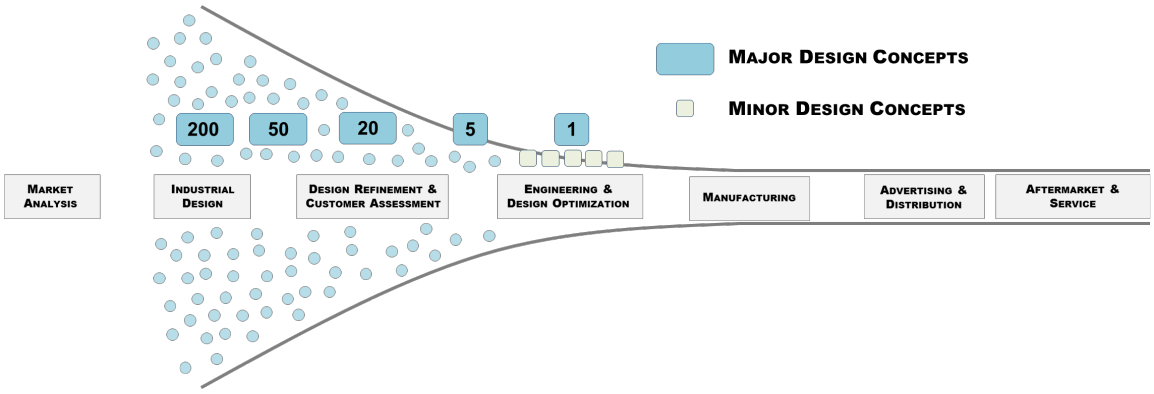


Figure 1.3: Depiction of the design process from the enterprise standpoint laid out in chronological order. Note that while general for many complex engineering products, this design process was recorded from interviews with practicing design executives at a major automotive manufacturer (Manoogian II, 2013; Hartley, 1996a), and may not generalize to all product or service designs. In particular, the partitioning or even existence of various design process steps may be different, as well as the number of major and minor design concepts. Further, note that while technically all major and minor design concepts are unique within the design space, we make the distinction between competing design concepts that are very far apart (major) and those that are small perturbations around a baseline design concept (minor). This distinction assumes some notion of distance within the design space.

Formalizing these concepts, the framework of design as a decision-making process has become the currently dominant paradigm in the way design is mathematically represented within the design research community (Hazelrigg, 1998; Chen *et al.*, 2013; Papalambros, 2002; Wassenaar & Chen, 2001). The design process is conceptualized as a number of sequential or parallel design decisions made by any number of decision makers at all levels of a hierarchy within the enterprise. Example decisions may be, "which emerging market segment is most suitable for new product development," or

"should we use a series, parallel, or power-split hybrid electric architecture," to "what ratio of thread to pitch should this grade 8 bolt be?"

In this dissertation, we use the term 'designer' to mean 'decision-maker' within the decision-based design framework. Designer is thus a loose term, and could be any number of various people involved in the design process, as long as they are associated with a given design task. Moreover, following the definition of crowdsourcing in Section 1.2.1, the designer for a given design task may use a crowdsourcing system made up of a crowd of other designers for other design tasks, rather than unknown people distributed across the internet. In other words, the term 'designer' is context dependent on the design task itself. We will note later that one of the most practical uses of crowdsourcing within a large complex engineering systems enterprise is indeed to use "internal crowdsourcing" to circumvent siloing and communication issues throughout the enterprise.

Subjective and Objective Decisions We establish and important distinction between *subjective design decisions* and *objective design decisions*. Subjective design decisions are those that depend on opinion or preferences within a given group of people, whereas objective design decisions have a "true score" regardless of who is asked. For example, the most preferred aesthetic styling of a next generation entry-level luxury vehicle depends greatly on who is asked, whereas the weight of the vehicle has a true measurable value regardless of who is asked.

The distinction between subjective and objective decisions leads to the corresponding type of human input required–preference relations for subjective design decisions, evaluations for objective design decisions. Likewise, this distinction also leads to the language we use to describe members of the crowd. If the designer and subsequent design task are given a subjective design decision, then members in the crowd who are queried for their input regarding this decision are called *customers*. Likewise, if

the designer and subsequent design task are given a objective design decision, then members in the crowd who are queried for their input regarding this decision are called *evaluators*.

Crowdsourcing is useful for both types of design decisions. Subjective design decisions, by definition, need a group of potential customers. For objective design decisions, oftentimes we require human evaluation for design decisions that cannot be feasibly or easily deduced through experimentation or computational simulation. This dissertation will use two recurring case studies, bracket topology optimization for objective evaluation and passenger vehicle aesthetic styling for subjective preferences, as benchmarks for our crowd aggregation models.

1.2.3 The Promise of Crowdsourcing for Making "Good" and Catching "Bad" Decisions

> "Quia parvus error in principio magnus est in fine, secundum philosophum ... (A small error at the outset can lead to great errors in the final conclusions, as the Philosopher says ...)." - Thomas de Aquinas, 1255

A well-known finding within design research is that the cost of making changes to the design increases monotonically and precipitously as the design reaches further stages of refinement during the design process. In other words, when the design concept(s) are just sketches, it is virtually free to make changes, while late-stage manufacturing changes or even changes after the final design embodiment lead to significant costs to the enterprise (e.g., recalls for defects after the customer has already purchased the design).

If we couple the notion of potentially crippling costs regarding early versus late-stage design changes, with the notion that product development lead times are ever shrinking given increased competition across the globalized world (Bloebaum *et al.*, 2012), we have a recipe for disaster if the enterprise is not able to make "good" design

decisions and catch "bad" design decisions in a timely manner during the product development process. An extreme example is that of the U.S. Department of Defense, which in a single year has on average $296 Billion and 22 months of cost and time overruns, respectively (Francis *et al.*, 2010; Oehmen & Seering, 2011).

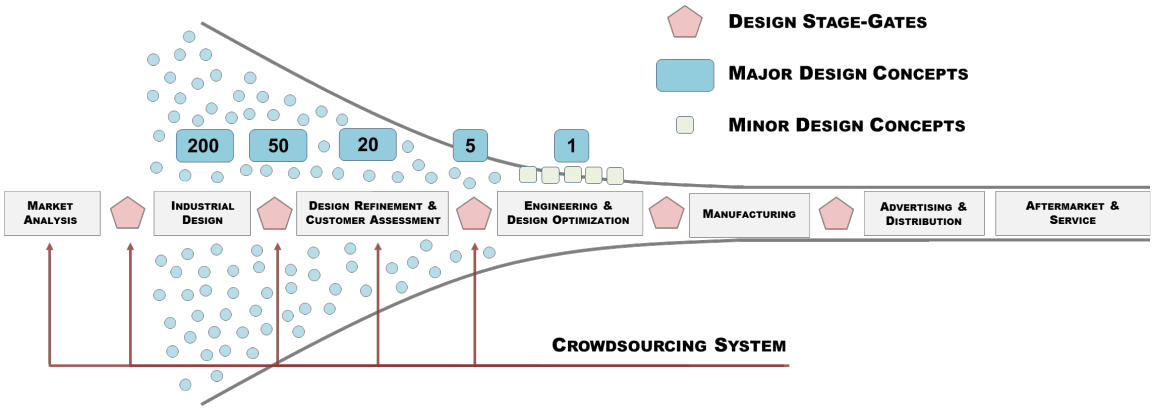


Figure 1.4: Depiction of design process augmented with crowdsourcing system to help designers make better decisions, particularly during design stage-gates.

These cost and time overruns have not gone unnoticed, and a number of research findings within the product innovation and corporate management academic literatures have shown that with ever more globalized markets and enterprises competing on the world stage, it is becoming increasingly imperative to incorporate design innovation to stay market competitive. The banner that these methods often go under is "open innovation," with crowdsourcing being one such method within this greater umbrella. In particular, there is potential for crowdsourcing to augment the designer's decision making at critical junctions during the design process termed "stage-gates" (Cooper, 1990; Hauser *et al.*, 2006). Specifically, before letting a design concept move on to later stages of the design process, a designer may seek relevant information distributed over a crowd (e.g., which of these 5 related concept designs is too conservative or too innovative for the market (Manoogian II, 2013)). We depict how crowdsourcing may improve "good" decisions and catch "bad" decisions during stage-gates in the

design process in Figure 1.4.

While it is difficult to say with certainty the number of enterprises engaged in crowdsourcing practices—no doubt unaccommodated by our practical definition of crowdsourcing, which does not explicitly demarcate between crowdsourcing, outsourcing, and simply asking people for evaluations—cursory analysis suggests there are over 2000 active websites related to crowdsourcing *crowdsourcing.org* (2014). Further, a number of academic publications and industrial white papers give some indication of the uses of crowdsourcing within industry, government agencies, and academia. We show in Figure 1.5 a small subset of enterprises engaged in crowdsourcing, chosen as their experiences have been analyzed by academic researchers or have helped drive adoption of the term crowdsourcing. One of the earliest successful uses of crowdsourcing by a large enterprise (though it was not called crowdsourcing at the time) is that of IBM's InnovationJam, in which 46,000 ideas, submitted over two periods of 72-hours, resulted in 10 new businesses and $100 million in funding (Bjelland & Wood, 2008).

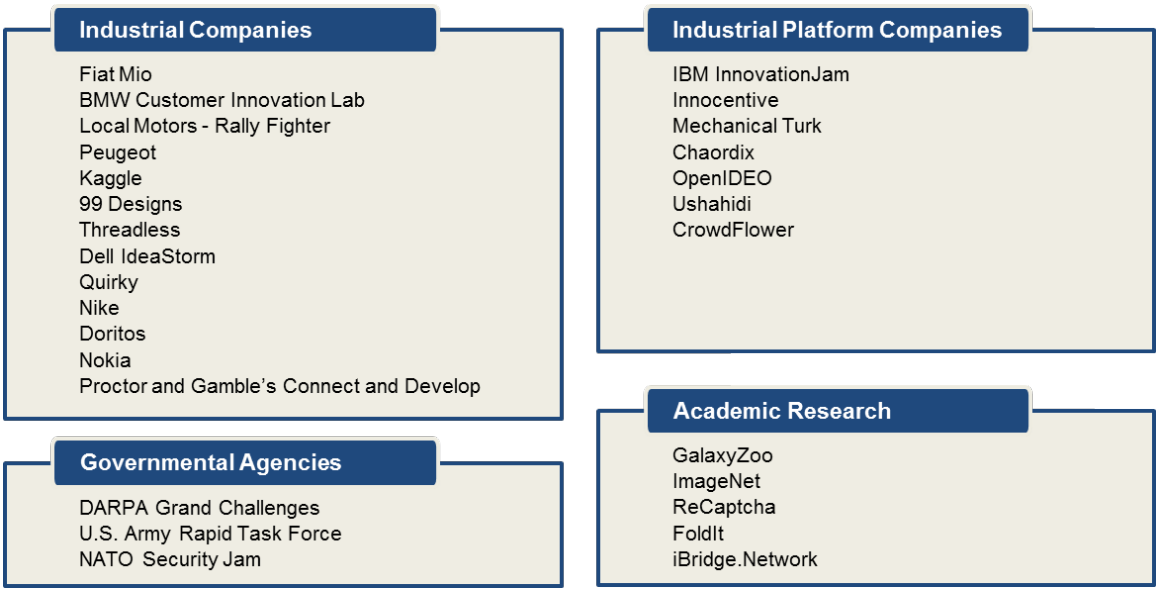


Figure 1.5: Selected subset of enterprises spanning industry, governmental agencies, and academia, engaged in crowdsourcing. This selection was made to cover enterprises that have had recurring academic and media coverage, as well as a diversity of enterprises exhibiting both successes and failures of crowdsourcing.

1.3 Quantitative Crowd Aggregation Models

We now introduce quantitative modeling techniques for aggregating crowd input, be it subjective preferences or objective evaluations, in order to parallel the qualitative motivation for crowdsourcing systems for the designer within the enterprise given in Section 1.2.3.

We begin with quantitative models of the decision-making process for a single evaluator or customer. The discussion starts axiomatically, listing the deficiencies of axiomatic approaches for real engineering design problems, and justifying the use of statistical approaches. Next, we examine multiple evaluators or customers constituting a crowd. We will see that the problem of crowd aggregation has been formally

studied for centuries, particularly in the area of voting theory, yet still leaves open challenges for engineering design leading to a number of research gaps later detailed in Section 1.4.

1.3.1 Single Evaluator Models

We begin with definitions of the customers or evaluators (note that we use these terms interchangeably at this point), designs, as well as preferences or evaluations between the designs by the customer or evaluator.

A single customer (or evaluator) $\mathbf{x}_c^{(j)} \in \mathcal{X}_c \subseteq \mathbb{R}^{M_c}, j = \{1, \dots C\}$ is represented by a vector of variables (e.g., age, gender, income) in an M_c-dimensional space. Similarly, a single design $\mathbf{x}_d^{(p)} \in \mathcal{X}_d \subseteq \mathbb{R}^{M_d}, p = \{1, \dots D\}$ is represented by a vector of variables (e.g., geometric shape, color, price) in an M_d-dimensional space. We use the superscript indices in parentheses, (j) for customers and (p) for designs, to denote the j-th customer or p-th design known to the designer or customer, whichever is later appropriate. Note that for practicality, we always assume these are finite or countably infinite sets.

As we introduced earlier, the designer wishes to improve her design decision making, and thus asks a *design task* to the customer or evaluator. For example, a subjective design decision may be: "Which of these concept passenger vehicle designs are you most likely to purchase?" Similarly, an objective design decision may be: "Which of these concept military vehicle designs has the most situational awareness?" We assume that, for this given design task, the customer or evaluator has a strict, linear ordering of designs (i.e., satisfying completeness, reflexive/antisymmetric, and transitive relations). Formally, customer $\mathbf{x}_c^{(j)}$ prefers or evaluates $\mathbf{x}_d^{(p)} \succ \mathbf{x}_d^{(q)} \dots \succ \mathbf{x}_d^{(r)}$ for the entire set of designs $p, q, r = \{1, \dots D\}$, in which $\succ$ denotes strictly prefers or strictly evaluates. We may equivalently associate a single permutation or ranking $\tau^{(j)} \in \mathcal{R}(\mathcal{X}_d)$ of the indices of set of designs $1, \dots, D$ with a given customer j,

corresponding to her ordering of designs, and in which $\mathcal{R}(\mathcal{X}_d)$ is the set of all linear ordering of designs. It is now up to the designer to deduce this ranking of designs to improve her decision-making during the design process. The designer would like to capture this ranking in the form of a mathematical function, one that inputs designs (or some subset of the designs) and outputs the corresponding correct ordering for the given customer.

This is a mature area of study, termed utility theory, as originally motivated by the Marquis de Condorcet (Condorcet, 1785) and formalized by von Neumann and Morgenstern (Von Neumann & Morgenstern, 2007). In short, utility theory is an approach to convert a set of designs to a set of numbers, while still preserving ordering properties. Formally, there exists a utility model $u^{(j)} : \mathcal{X}_d \times \mathcal{X}_c \to \mathcal{Y}$ pairing each design with a number (typically the reals, i.e., $\mathcal{Y} = \mathbb{R}$) for a given customer j, for the assumptions on finite or uncountably infinite linear orderings defined above (ord, n.d.). With knowledge of this utility model, a customer's ranking $\mathbf{x}_d^{(p)} \succ \mathbf{x}_d^{(q)} \ldots \succ \mathbf{x}_d^{(r)}$ may be deduced by ranking an associated set of values of corresponding designs as mapped by $u^{(j)}(\mathbf{x}_d^{(p)}, \mathbf{x}_c^{(j)}) > u^{(j)}(\mathbf{x}_d^{(q)}, \mathbf{x}_c^{(j)}) \ldots > u^{(j)}(\mathbf{x}_d^{(r)}, \mathbf{x}_c^{(j)})$. More generally, we aim to find the utility function $U^{(j)} : (\mathcal{X}_d)^D \times \mathcal{X}_c \to (\mathcal{Y})^D$ from which we can sort $\mathbf{y}$ to obtain the corresponding ranked permutation of designs $\tau^{(j)}$ for the customer j.

Problems with Axiomatic Methods for Engineering Design

Obtaining this utility model, relating a number for a given customer with a given design, is nontrivial for a variety of reasons. The first major problem with axiomatic approaches is sheer size of possible preference or evaluation comparisons possible for a given set of designs. Even in the case of perfect preference or evaluation responses by the customer or evaluator, asking a customer or evaluator to rank an entire set of designs is unreasonable due to the effectiveness of human judgment between large sets to rank versus smaller sets of designs to rank. Accordingly, instead of directly

estimating the ranking function, often a small subset of the full set of designs is asked for ranking, many times just 2 at a time for a binary preference relation (Hüllermeier *et al.*, 2008; Herbrich *et al.*, 1998). Decomposition of the full ranking to consistent binary pairs may be done to create the set $\mathcal{S}^{(j)} = \{(\mathbf{x}_d^{(p)}, \mathbf{x}_d^{(q)}), \text{sign}(y^{(p)}, y^{(q)})\} \subset \mathcal{X}_d \times \mathcal{X}_d \times \{-1, +1\}$. For full enumeration of the complete ranking, this results in $D(D-1)/2$ design pairs for evaluation, again an unreasonable request, this time due to the number of designs evaluations. As a result, we often ask for only a small subset of design evaluations, resulting in only a known partial ordering for each customer or evaluator.

The second major problem with axiomatic approaches is the irrationality of human behavior. Even if the customer or evaluator did not fatigue, and we were able to ask for a complete ranking of evaluations or preferences, human behavior is often far from rational. While there are entire research fields dedicated to irrational economic behavior, we make note now on results showing evaluation and preference inconsistencies in engineering design (MacDonald *et al.*, 2009). In particular, human behavior often exhibits violations of reflexivity (e.g., answering the same preference or evaluation differently at separate times; violations of symmetry (e.g., "Do you see yourself driving a Ferrari or Prius?" vs. "Do you see yourself driving a Prius or Ferrari?"); and violations of transitivity (e.g., $A \succ B, B \succ C \Leftrightarrow A \succ C$) (Simon, 1956).

The third major problem with axiomatic approaches, and perhaps the most important in terms of interpreting results for later actionable design decisions, is that we simply do not know all the relevant customer variables, design variables, their respective representations, and their deterministic dependency (i.e., functional form). For example, given a design decision regarding aesthetic styling of a passenger car, we may be missing relevant customer or design variables (e.g., the customer prefers round body shapes that elicit eco-friendliness (Reid *et al.*, 2010b)).

Practicality: From the Axiomatic Approach to the Probabilistic Approach

Given the number of challenges with implementing axiomatic deterministic approaches to asses utility models, we turn to stochastic approaches. In other words, our originally deterministic utility function is now treated stochastically, and we turn to statistical estimation procedures. Formally, we move from our original deterministic model of utility for customer or evaluator j:

$$u^{(j)}(\mathbf{x}_d, \mathbf{x}_c^{(j)}) = y \tag{1.1}$$

to a probabilistic model of utility for customer or evaluator j,

$$p^{(j)}(y|\mathbf{x}_d, \mathbf{x}_c^{(j)}, \Omega). \tag{1.2}$$

We use Ω to capture all the factors other than known design variables $\mathbf{x}_d$ and known customer variables $\mathbf{x}_c$ that affect customer j's response during the crowdsourcing process. For example, given a design task (e.g., rate the aesthetic styling of this vehicle concept according to its aggressiveness compared to the leading market competitor), we may ask the "same" customer (same age, same gender, etc.) and get wildly varying answers. This is understandable, as we will essentially never have full knowledge of these factors (e.g., the customer just had some bad food poisoning and everything seems aggressive). Enumerating these factors contributing to uncertainty (i.e., what is in Ω), we may have:

- Known Unknowns in Ω
  - Customer attributes $\mathbf{a}_c(\mathbf{x}_c)$ - Factors that are known to the designer, are functions of the customer variables themselves, but are hard to deduce explicitly, e.g., "environmentally-conscious customer"
  - Design attributes $\mathbf{a}_d(\mathbf{x}_d)$ - Factors that are known to the designer, are func-

tions of the design variables themselves, but are hard to deduce explicitly, e.g., "eco-friendly design."

  - Model parameters θ - Factors that govern the probability distribution for a model from a given model class $\hat{\mathcal{M}}$, e.g., mean and variance of a Gaussian. This is equivalent to the analogous fully deterministic model, e.g., coefficients $\mathbf{A}$ of a linear model $\mathbf{Ax} = \mathbf{b}$.

- Unknown Unknowns in Ω

  - Customer features $\mathbf{h}_c$ - Factors that are not known to the designer and are functions of the known customer variables and other unknown variables, e.g., "truck-owning liberal voter"
  - Design features $\mathbf{h}_d$ - Factors that are not known to the designer and are functions of the known design variables and other unknown variables, e.g., "Chevy Nova did not sell well in Mexico since 'No va' means 'no go' in Spanish"
  - True model family $\mathcal{M}^*$ - The true relationship among all known knowns, unknown knowns, and unknown unknowns. As we do not know this, we always make an assumption on the model class $\hat{\mathcal{M}}$ and hope we overlap, and if subsume, hope the model class itself is not too large to make statistical estimation inefficient.

1.3.2 Crowd Aggregation Models

As we have seen in Section 1.3.1, we may mathematically model the preferences or evaluations of a single customer or evaluator using utility theory, specifically via a probabilistic approach to account for real-world human behavior as well as practicality with data collection. Recall that, as our main goal with a crowdsourcing system is to support the designer and help her make better decisions during the design process,

our job now is to combine the input from all crowd members into a manageable summary useful to the designer via relevant information regarding the design decision in question. This manageable summary will be in the form of an aggregation of the customers or evaluators constituting the crowd, leading to clusters of similar evaluations or clusters of market segments. We will see that, in the naïve case, and indeed the most commonly used case in design and marketing research, we only have one cluster–an assumption that will be tested later in this dissertation.

We define the crowd $\mathbf{X}_c = \{\mathbf{x}_c^{(j)}\}_{j=1}^{C}$ as the entire set of evaluators in our crowdsourcing system as seen by the designer. The goal of a crowdsourcing system is to collect and aggregate the preferences or evaluations of the crowd, in the perfect case represented as a a collection of heterogeneous linear orderings $T = \{\tau^{(j)}\}_{j=1}^{C}, \tau^{(j)} \in \mathcal{R}(\mathcal{X}_d)$ to a set of K preference or evaluation rankings that help the designer make a decision. Formally, the crowdsourcing system acts as a function $F : \mathcal{R}(\mathcal{X}_d)^C \rightarrow \mathcal{R}(\mathcal{X}_d)^K$. In this manner, the K may be interpreted as clusters of similar preferences or evaluations, and may be all customers' rankings or just a single ranking, i.e., $K = D$ or $K = 1$, respectively. Similar to the single evaluator or customer case, we will assume a utility function exists for each customer or evaluator, which will be captured as $\mathbf{Y}$, a $C \times D$ matrix of individual evaluator utilities.

Similar to the single customer or evaluator case discussed in Section 1.3.1, deterministically deducing the values of $\mathbf{Y}$ is challenging for many engineering design problems due to incomplete preference or evaluation data as well due to human behavioral issues. Accordingly, we again move to the probabilistic approach for aggregation of customer preference or evaluator orderings $\tau^{(j)}$ as mapped via their corresponding utility function $u^{(j)}$, and treat $\mathbf{Y}$ as a random matrix,

$$p(\mathbf{Y}|\mathbf{X}_c, \mathbf{X}_d, \Omega), \tag{1.3}$$

such that sorting each row $\mathbf{y}^{(j)}$ results in the evaluation ranking or preference ranking

$\tau^{(j)}$ over all designs. In other words, each element of the matrix $\mathbf{Y}$ is an instance of Equation (1.2). Like its analogous single evaluator formulation, Equation (1.3) is a general probabilistic equation for the crowd's utility function, one which generalizes the five research chapters in this dissertation. At the same time, this equation again is uninformative alone, as all the differences between various models are due to various random factors within Ω; for example, Ω may include random variables which tie together all evaluators at the crowd level.

What is different now, compared to the single evaluator case, is that we have access to a number of other customers or evaluators whose preferences or evaluations we may leverage. In this view, there are two major approaches amongst crowdsourcing aggregation methods: (1) Content-based aggregation, in which we are explicitly using known customer $\mathbf{X}_c$ and design variables $\mathbf{X}_d$ when inferring utility values $\mathbf{Y}$, or (2) collaborative filtering, in which we only use values in $\mathbf{Y}$ as computed from preferences or evaluations, though sometimes implicitly inferring the unknown customer $\mathbf{X}_c$ and design variables $\mathbf{X}_d$.

Collaborative filtering tends to perform better purely from a prediction accuracy standpoint, as it abstracts away the customer and design variables and instead uses unknown but relevant features. This is not a paradox, in that we are not doing better from a prediction accuracy standpoint with less information (i.e., no knowledge of customer $\mathbf{X}_c$ and design variables $\mathbf{X}_d$); instead, this is due to the type of prediction problems collaborative filtering is applied to. Specifically, collaborative filtering works well with a large number of previously observed data indicating preferences or evaluations for the same customer. Examples include music recommendation or search engine queries, in which even for a single customer or evaluator, a large number of previous "upvotes" or "downvotes" or "chosen web link out of a ranked search" is available. Collaborative filtering is able to look at other customers with similar patterns of "upvotes" and "downvotes" to predict overall utility values $\mathbf{Y}$ and ultimately

ranked permutation $\tau^{(j)}$ of designs (in this case songs), without knowing anything about the customer $\mathbf{X}_c$ (e.g., age, gender, culture) or design $\mathbf{X}_d$ (e.g., melodic patterns, drum beat styles). In this manner, collaborative filtering acts as a "black box" preference or evaluation function, which in general has limited general usefulness to the designer.

Content-based preference or evaluation aggregation directly uses customer $\mathbf{X}_c$ and design variables $\mathbf{X}_d$, and consequently, has major advantages for designers due to the fact that it is often much easier to interpret as the designer has direct access to the underlying variables that affect preferences or evaluations. For example, the designer may see that environmentally-conscious customers may strongly prefer vehicles that are 4-cylinder hybrids less than 1.5 liters in displacement, and which have slow acceleration. By assessing these preferences, and in particular the customer or design variables that most influence preferences or evaluations, the designer is more able to ultimately make actionable design decisions that positively affect the end design.

Another major reason we are interested in crowdsourcing systems built using content-based preference or evaluation aggregation is because many real engineering design problems simply do not have previous "upvotes" and "downvotes"–in other words, content-based preference and evaluation aggregation is oftentimes the only approach that works for particular engineering design tasks. For example, the automobile purchase decision is one of the most significant personal decisions a family must make regarding both quality of life as well as personal finances. This decision only occurs on average every 6 years, leading to the average age of vehicles on the U.S. road at 11 years old. Moreover, one's decision 6 years ago likely is much less relevant than the current set of variables governing ones present life (e.g., number of children, income, location of residence). Contrast this decision with that of search queries on an internet search engine, or upvoting and downvoting songs you enjoy on an online music streaming service, in which a large corpus of your previous decisions

is captured, decisions that are relevant due to recency and relatively-static music preferences over this short period.

Accordingly, while we use both collaborative-filtering-based approaches (Chapters 2, 4, and 6) and content-based preference and evaluation aggregation (Chapters 3, 5), in the end we are most interested in crowdsourcing systems that use the latter due to its usefulness for designers and resulting actionable design decisions. This is fortuitous, as though we are currently in a period of increasingly large dataset sizes (thus increasing the performance of both approaches), content-based aggregation has perhaps more opportunity for research advances compared with collaborative filtering. This is due to (1) the task-specificity of content-based aggregation, and (2) the relatively recent advances in the number and sophistication of content-based aggregation of preferences and evaluations. We will discuss these notions in detail throughout the rest of this dissertation, particularly in the Chapters 3, 5, and 6.

Name	y	$\mathbf{X}_c, \mathbf{X}_d$	Ω	$\hat{\mathcal{M}}$	**Reference**
Item Response Theory	$\{0, 1\}$	-	Expertise, Difficulty, Scaling, Bias	Logit	(Lord, 1952)
Rasch Models	$\{0, 1\}$	-	Expertise, Difficulty	Logit	(Rasch, 1966, 1960/1980)
Placket-Luce	$\tau^{(j)}$	-	-	Rank-Ordered Logit	(Plackett, 1975; Luce, 1959)
Thurstone Model	$\tau^{(j)}$	-	-	Rank-Ordered Logit, Gaussian	(Thurstone, 1927)
Conjoint Analysis, Discrete Choice	$\{0, 1\}$	Designs	-	Logit	(McFadden & others, 1973)
Dawid-Skene	$\{0, 1\}$	Designs	Expertise, Difficulty via Confusion Matrix	Logit	(Dawid & Skene, 1979)

Table 1.1: Seminal crowd aggregation models from the research disciplines of psychometrics, econometrics, social welfare models. Note that we only give the basic form of these models, as most of these models have extensions to include additional terms.

Name	y	$\mathbf{X}_c, \mathbf{X}_d$	Ω	$\hat{\mathcal{M}}$	**Reference**
Bayesian Truth Serum	$\{0, 1\}$	-	Common Prior on "Expertise"	Undefined	(Prelec, 2004a)
DARE	$\{0, 1\}$	-	Expertise Clusters, Difficulty, Expertise Scaling	Gaussian	(Bachrach *et al.*, 2012b)
Instrumenting Crowd	$\{0, 1\}$	Customers	-	Decision Tree	(Rzeszotarski & Kittur, 2011)
Raykar	$\{0, 1\}$	Designs	Expertise, Difficulty via Confusion Matrix	Logit	(Raykar *et al.*, 2009)
GLAD	$\{0, 1\}$	-	Expertise, Difficulty, Adversaries	Logit	(Welinder *et al.*, 2010a; Whitehill *et al.*, 2009a)

Table 1.2: Modern crowd aggregation models from statistics and machine learning. Note that we only give the basic form of these models, as most of these models have extensions to include additional terms.

Seminal and Modern Crowd Aggregation Models

While there are hundreds of models of crowd aggregation that all fall under a similar umbrella, Tables 1.1 and Table 1.2 give an overview of selected seminal works from psychometrics, econometrics, and social welfare models, and relate these models to the general crowd aggregation model given in Equation (1.3). Most of these

models have seen extension to different evaluation types (e.g., rating, choice, partial ranking), extensions to fully probabilistic Bayesian or empirical Bayesian formulations, methods of parameter estimation (e.g., Markov Chain Monte Carlo (MCMC), various loss functions; however, we only give their most basic form as introduced in their seminal works. We similarly include modern crowd aggregation models from statistics, machine learning, and the crowdsourcing community. Given that most of these recent models are repeats or extensions of historic models, we choose to list modern models that are qualitatively different from each other.

1.4 Research Gap and Dissertation Contributions

As we have seen in Section 1.2.3, there is much qualitative justification for a crowdsourcing system to improve "good" and to catch "bad" design decisions during the early-stage design process in an effort to mitigate cost and time overruns for the enterprise. To build these crowdsourcing systems, we have discussed quantitative models for crowd aggregation in Section 1.3.

While success has been shown with crowdsourcing systems within certain niches in which a single expert designer is fully capable of outputting a complete design embodiment, such as graphic design and image annotation, there is less reported success with companies engaged in "complex" design problems. In fact, current business case studies and academic literature report that there are significantly more unsuccessful uses of crowdsourcing by enterprises (Chiu *et al.*, 2014).

To gain qualitative insight to successful crowdsourcing for complex designs include, we may look at the following examples: (1) Boeing's concurrent engineering processes involve creating cross-functional teams of experts from a number of relevant disciplines to evaluate design concepts at stage-gate reviews (Klein *et al.*, 2006); (2) IBM's InnovationJam used an expert crowd of 50 internally selected executives to evaluate innovative design concepts (Bjelland & Wood, 2008; Blohm *et al.*, 2013); and (3)

Fiat's Mio used its internal engineering design team to evaluate crowdsourced design concepts (Celaschi *et al.*, 2011a).

The observation here is that these companies have shown success with crowdsourcing through the use of expert evaluators or lead uses as an ad hoc filtering mechanism for the crowd itself, certainly not the systematic crowd aggregation methods used in simple problems such as image annotation. In other words, a major reason crowdsourcing is often unsuccessful for complex designs is in the filtering of low-quality submissions–there is often too little "signal" to "noise" in the crowd for complex design problems (Peisl *et al.*, 2014). Instead, these successes are often only the result of a small subset of motivated expert evaluators or lead users in a market (Chiu *et al.*, 2014; Dahlander & Gann, 2010).

The academic literature thus suggests that there are major practical limitations to the usefulness of current crowdsourcing processes within companies engaged in complex engineering design. In particular, the large number of low-quality submissions results in low "signal-to-noise," as well as the tendency for successful crowdsourcing to be the result of just a small subset of the overall crowd with appropriate *expertise*, leads to heterogeneity of expertise across various evaluators or customers in the crowd.

Parallel to these qualitative findings, the academic literature is rife with mathematical models of crowd aggregation for simple design tasks. As was shown in Tables 1.1 and 1.2, crowdsourced aggregation models typically do not include terms explicitly accounting for: (1) Variables representing the evaluators or customers themselves, and (2) heterogeneity of expertise or preferences. If these models do include expertise, they are applied to "simple" tasks in which a majority of the crowd is able to perform the task, and in which "expertise" maps to attention or fatigue.

Combining the qualitative justifications for crowdsourcing systems for enterprises engaged in complex engineering design, with the lack of quantitative models tailored

to this case, the research gap pursued in this dissertation is:

> To quantitatively investigate why the heterogeneity of evaluator expertise and customer preferences has led to unsuccessful crowdsourcing systems, and to develop crowd aggregation models of both objective and subjective design decisions that mitigate these issues and lead to practical crowdsourcing systems for engineering design.

Our main research contribution in this dissertation, which is summarized in greater detail in Section 7.2, is a quantitative study on the above gap across the spectrum of objective design decisions to subjective design decisions in complex engineering design. This systematic study provides quantitative understanding why crowdsourcing systems fail for both objective design decisions and subjective design decisions due to the heterogeneity of evaluator expertise or customer preferences.

For objective design decisions, heterogeneity in expertise results in "consistently wrong clusters" of evaluators that are statistically impossible to find and filter out, thus washing out the combined crowd evaluation. For subjective design decisions, heterogeneity in preferences results in a number of different optimal designs, which get missed by traditional design preference models.

We develop a number of probabilistic crowd aggregation models that capture and thus mitigate heterogeneity in the crowd, and contribute to practical crowdsourcing systems for engineering design.

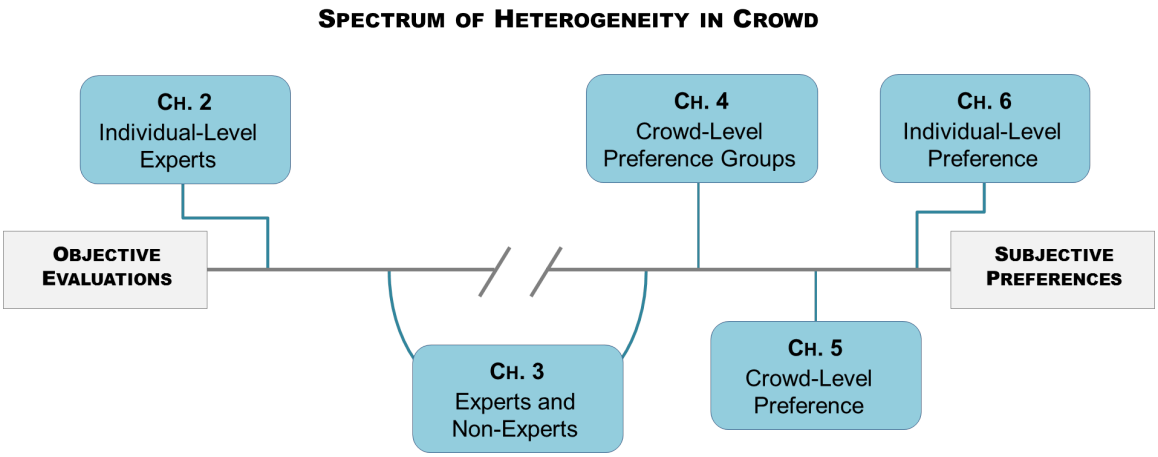


Figure 1.6: Spectrum of heterogeneity for a given objective or subjective design decision. The left-hand end is the case of only a very sparse minority of the crowd having enough expertise for the given design task. In between both extremes are various levels of expertise needed for a given design task. The right-hand end is the case in which by definition no expertise is needed due to individual-level preferences.

1.5 Dissertation Overview

The rest of this dissertation consists of five chapters of research findings conceptually organized as shown in Figure 1.6. The dissertation is organized by moving from right to left on this "spectrum of expertise needed" for a given design decision. In other words, we start with the case of a design decision that requires a very high level of expertise to correctly evaluate, and move to the case in which the design decision requires no expertise by definition.

Chapter 2 models the crowd aggregation process by including evaluator expertise and design difficulty, in a manner qualitatively similar to many of the models in Tables 1.1 and 1.2. We show that these models fail even for "simple" engineering design tasks due to the relative sparsity of expertise in the crowd. Chapter 3

attempts to find discriminative information to identify experts in a crowd of experts and non-experts, such that non-experts may be filtered out for a better crowd aggregation. We show that many evaluator variables such as demographics, reaction times, and benchmark mechanical skills tests fail to predict expertise; instead, evaluation expertise is predicted by performance on an"easy" version of the actual "hard" engineering design task. Chapter 4 moves to the case in which the design task consists of multiple clusters of "expertise," specifically brand recognition, in which we now filter out non-experts. Chapter 5 examines the case of a single crowd-level preference, namely, the visual realism of designs generated using an algorithm. Chapter 6 finally deals with the case in which no expertise is needed for the design task by definition, in which the heterogeneity of every customer is modeled as everyone is an expert of their own individual-level preferences. Chapter 7 concludes by summarizing key takeaways, contributions to engineering design, and opportunities for future work.

CHAPTER II

Why does Crowdsourcing Fail for Objective Evaluations?

2.1 Context: Do current crowdsourcing aggregation models work for engineering design?

Suppose we wish to evaluate a set of vehicle design concepts with respect to attributes that have objective answers. For many of these objective attributes, the "true score" may be determined using detailed physics-based simulations, such as finite-element analysis to evaluate crashworthiness or human mobility modeling to evaluate ergonomics; however, for some objective attributes such as maintainability, physics-based simulation is difficult or not possible at all. Instead, these objective attributes require human input for accurate evaluation.

To obtain evaluations over these objective attributes, one may ask a number of specialists to evaluate the set of vehicle design concepts. This assumes the requisite expertise is imbued within this group of specialists. Oftentimes though, the expertise to make a comprehensive evaluation is instead scattered over the "collective intelligence" of a much larger crowd of people with diverse backgrounds (Hong & Page, 2004b).

Crowdsourced evaluation, or the delegation of an evaluation task to a large and

possibly unknown group of people through an open call (Estellés-Arolas & González-Ladrón-de Guevara, 2012; Gerth *et al.*, 2012), is a promising approach to obtain such design evaluations. Crowdsourced evaluation draws from the pioneering works of online communities, such as Wikipedia, which have shown that accuracy and comprehensiveness are possible in a large crowdsourced setting requiring expertise. Although crowdsourcing has seen recent success in both academic studies (Kittur *et al.*, 2008) and industry applications (Von Ahn *et al.*, 2008; Warnaar *et al.*, 2012), there are limited reference materials on the use of crowdsourced evaluation for engineering design.

In this chapter, we explore how the expertise of evaluators in the crowd affects crowdsourced evaluation for engineering design, where expertise is defined as the probability that a evaluator gives an evaluation close to the design's true score. The choice of exploring expertise comes from an important lesson in managing successful online community efforts, namely, the need to implement a systematic method of filtering "signal" from "noise" (Ipeirotis & Paritosh, 2011). In a crowdsourced evaluation process, this manifests itself as a need of screening good evaluations from bad evaluations, in particular when we are given a heterogeneous crowd made up of a mixture of expert and non-expert evaluators. In this case, averaging evaluations from all participants with equal weight will reduce the accuracy of the crowd's combined evaluation, also called the *crowd consensus* (Sheshadri & Lease, 2013a), due to incorrect design evaluations from low-expertise evaluators. Accordingly, a desirable goal is to identify the experts from the rest of the crowd, thus allowing a more accurate crowd consensus by giving their evaluations more weight.

With this goal in mind, we developed and benchmarked a crowd consensus model of the crowdsourced evaluation process using a Bayesian network that does not require prior knowledge of the true scores of the designs or the expertise of each evaluator in the crowd, yet still aims to estimate accurate design scores by identifying the experts

within the crowd and overweighting their evaluations. This statistical model links the expertise of evaluators in the crowd (i.e., knowledge or experience for the design being evaluated), the evaluation difficulty of each design (e.g., a detailed 3D model provides more information than a 2D sketch and may therefore be easier for an expert to evaluate accurately), and the true score of each of the designs. It must be noted that this model relies *only* on evaluations from the crowd; i.e., we do not explicitly measure expertise or difficulty; these variables are latent and only implicitly inferred.

This crowd consensus model rests on the key assumption that low-expertise evaluators are more likely to "guess," and are thus more likely to give random evaluations to designs. This assumption is modeled by defining an evaluation as a random variable centered at the true score of the design being evaluated (Nunnally & Bernstein, 2010). A graphical representation of the Bayesian network showing these relationships is given in Figure 2.2.

The performance of the Bayesian network crowd consensus model versus the baseline method of averaging evaluations is explored through two studies on the same "simple" engineering design evaluation task of rating the strength of a load-bearing bracket (Papalambros & Shea, 2005). First, we created simulated crowds to generate evaluations for a set of designs. These crowds had a homogeneous or heterogeneous expertise distribution, representing two cases that may be found in a human crowd. Second, we used a human crowd recruited from the crowdsourcing platform Amazon Mechanical Turk (Amazon, 2005), and performed a crowdsourced evaluation with the same crowd and task properties as in the simulation.

Our results show that we are *not* able to achieve a more accurate design evaluation using the crowd consensus model than just averaging all evaluations. Even for the simple engineering design evaluation task in this chapter, the modeling assumption that low-expertise evaluators guess more randomly was found not to hold. Upon further investigation, it was found that there exist numerous clusters of "consistently

wrong" evaluators that wash out the evaluations from the cluster of experts.

The main contribution of this chapter is this finding; namely, that crowdsourced evaluation can fail for even a simple engineering design evaluation task due to the expertise distribution of the crowd. Averaging already gives a low-accuracy estimate of design scores due to the large number of low-expertise evaluators, and a crowd consensus model relying *only* on information about evaluations may not be able to find the experts in the crowd. This chapter thus serves as justification for further research into methods of finding experts within crowds, particularly when they are shrouded by numerous clusters of consistently wrong non-experts.

The remainder of this chapter is organized as follows. Section 2.2 reviews relevant research within the engineering design, psychometrics, and crowdsourcing literature, as well as research motivations from industry. Section 2.3 presents the Bayesian network crowd consensus model and modeling assumptions. Section 2.3 details the statistical inference scheme of the Bayesian network. Section 2.4 describes the simulated crowds experiment and results. Section 2.5 describes the human crowd experiment and discusses its results. We conclude in Section 2.8 with implications of this work and opportunities for future research.

2.2 Related Work

Within the engineering design community, attention is being drawn to the use of crowdsourcing for informing design decisions (Van Horn *et al.*, 2012). Design preferences have been captured using crowdsourced data on social media sites (Tuarob & Tucker, 2013; Stone & Choi, 2013), as well as through more directed crowdsourced elicitation using online surveys for preference learning (Ren & Papalambros, 2012a; Ren *et al.*, 2013b). Our work differs from these works in that we focus on design evaluation with an objective answer, thus necessitating the estimation of evaluator expertise. Within design evaluation for objective attributes, recent research has used

crowdsourcing for idea evaluation (Kudrowitz & Wallace, 2013; Grace *et al.*, 2014) and creativity evaluation (Fuge *et al.*, 2013). There also exists much research studying the effect of a single decision maker versus crowd consensus decisions (Yang, 2010; Gurnani & Lewis, 2008). Our work is relevant to these research efforts in that we extend previous findings of the potential limitations of using the entire crowd for design evaluation.

Modeling the crowdsourced evaluation process exists in the literature extending at least back to Condorcet (de Caritat *et al.*, 1785), with foundational contributions from the psychometrics community under Item Response Theory (Lord, 1980) and Rasch Models (Rasch, 1960/1980). These models have been applied to standardized tests, with several extensions to include hierarchical structure (Oravecz *et al.*, 2013) similar to the crowd consensus model in this work. Additional foundational literature from econometrics includes "mechanism design" such as prediction markets and peer prediction (Miller *et al.*, 2005; Prelec, 2004b). For simplicity, we do not consider important findings and approaches from this econometrics literature, instead assuming all evaluators give truthful evaluations and are similarly incentivized by a fixed-sum payment.

More recently, the crowdsourcing community has developed numerous crowd consensus models capturing the expertise of evaluators in a crowdsourced evaluation process (Sheshadri & Lease, 2013a). Many of these models are qualitatively similar, with differences in the treatment of evaluator bias (Wauthier & Jordan, 2011; Bachrach *et al.*, 2012a; Welinder *et al.*, 2010b), form of the likelihood function (e.g., ordinal, ranking, binary) (Lakshminarayanan & Teh, 2013), extent to which the true score is known (Tang & Lease, 2011), and methods of scaling up to larger data sets (Welinder *et al.*, 2010b; Liu *et al.*, 2012). These models are most often applied to tasks that are "human easy, computer hard," such as image annotation (Whitehill *et al.*, 2009b; Welinder *et al.*, 2010b), planning and scheduling (Kim *et al.*, 2013),

and natural language processing (Snow *et al.*, 2008; Zaidan & Callison-Burch, 2011). Our research is also qualitatively similar to this literature, but with a key difference on the application to an engineering design task and the subsequent distribution of expertise in the crowd.

Specifically, many of these recent crowdsourced evaluation efforts are applied to tasks in which a majority of evaluators within the crowd have the expertise to give an accurate evaluation (e.g., does this image contain a 'duck'?) (Sheshadri & Lease, 2013a). As a result, either averaging or taking a majority vote of the crowd's evaluators is often already quite accurate (Sheng *et al.*, 2008a). For these cases, expertise may often represent the notion of task consistency and attentiveness, with low-expertise evaluators being more spammy or malicious (Welinder *et al.*, 2010b).

In contrast, many engineering design tasks may require expertise that only exists in a sparse minority of the crowd. This notion is supported by prior industrial applications of crowdsourced evaluation for engineering design. The Fiat Mio was a fully crowdsourced vehicle design concept, yet the large number of low-expertise submissions resulted in Fiat using its design and engineering teams as a filter without the use of algorithmic assistance (Celaschi *et al.*, 2011b). Local Motors Incorporated developed the Rally Fighter using a crowdsourced evaluation system similar to this research, but heavily weighted evaluations of the internal design team (Bommarito *et al.*, 2011). For these engineering design tasks, the notion of expertise may instead represent specialized knowledge and heuristics necessary to give an accurate evaluation.

2.3 A Bayesian Network Model for Crowd Aggregation

We introduce a crowd consensus model that statistically aggregates the evaluations from the set of evaluators using a Bayesian network to estimate the true design scores. More formally, let the crowdsourced evaluation contain D designs and P evaluators.

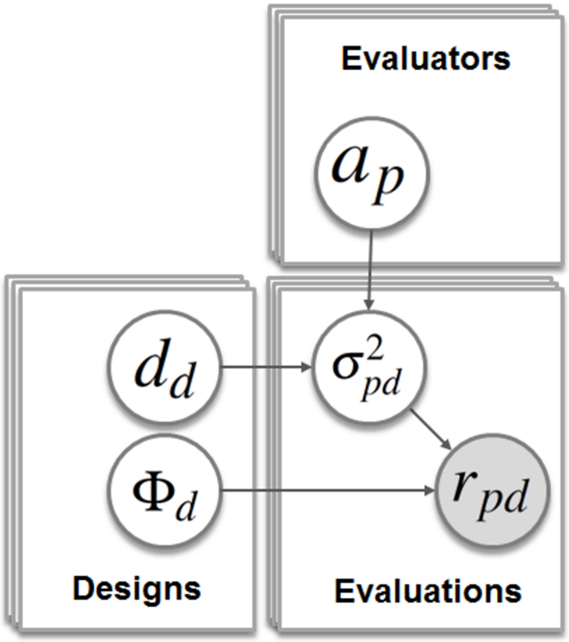


Figure 2.1: Graphical representation of the Bayesian network crowd consensus model. This model describes a crowd of evaluators making evaluations r_{pd} that have error from the true score Φ_d. Each evaluator has an expertise a_p and each design has an difficulty d_d. The gray shading on the evaluation r_{pd} denotes that it is the only observed data for this model.

We denote the true score of design d as $\Phi_d \in [0, 1]$, and the evaluation from evaluator p for design d as $\mathbf{R} = \{r_{pd}\}$ where $r_{pd} \in [0, 1]$. Each design d has an evaluation difficulty d_d, and each evaluator p has an evaluation expertise a_p.

The evaluation r_{pd} is modeled as a random variable following a truncated Gaussian distribution around the true performance score Φ_d as detailed by Eq. (2.1) and shown in Figure 2a.

$$r_{pd} \sim \text{Truncated-Gaussian}\left(\Phi_d, \sigma^2_{pd}\right), \quad r_{pd} \in [0, 1] \tag{2.1}$$

The variance of density σ^2_{pd} is interpreted as the error an evaluator makes when using his or her cognitive processes while evaluating the design, and is described by a random variable taking an Inverse-Gamma distribution:

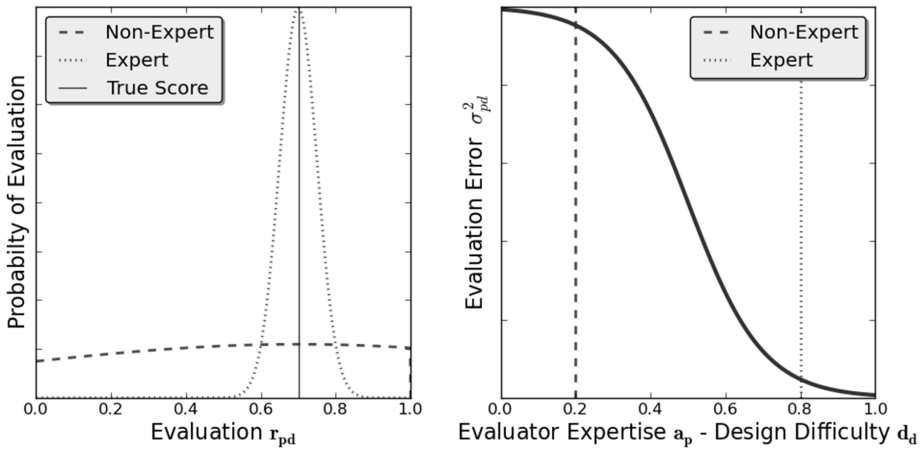


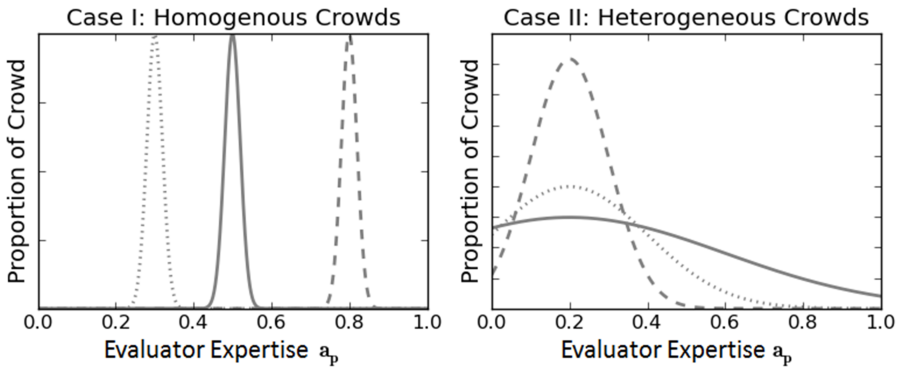
Figure 2.2: (a) Low evaluation expertise (dashed) relative to the design evaluation difficulty results in an almost uniform distribution of an evaluator's evaluation response, while high evaluation expertise (dotted) results in evaluators making evaluations closer to the true score. (b) An evaluator's evaluation error variance σ^2_{pd} as a function of that evaluator's expertise a_p given some fixed design difficulty d_d and crowd-level parameters θ and γ.

$$\sigma^2_{pd} \sim \text{Inverse-Gamma}\left(\alpha_{pd}, \beta_{pd}\right) \tag{2.2}$$

The average evaluation error for a given evaluator on a given design is a function of the evaluator's expertise a_p and the design's difficulty d_d. In addition, this function is sigmoid to capture the notion that there exists a threshold of necessary background knowledge to make an accurate evaluation. Figure 2b illustrates this function. We set the first requirement on the evaluator's error random variable using the expectation operator $\mathbb{E}$ in Eq. (2.3).

$$\mathbb{E}\left[\sigma^2_{pd}\right] = \frac{1}{1 + e^{\theta(d_d - a_p) - \gamma}} \tag{2.3}$$

The random variables θ and γ are introduced as model parameters to allow more flexibility in modeling evaluation tasks and are assumed to be the same for all evaluators and designs: A high value of the scale parameter θ will sharply bisect the crowd into good evaluators with negligible errors and bad evaluators that evaluate almost



Case	Type of Crowd	Varied Parameter	Figure	# Sim. Crowds
I	Homogeneous Crowd	Avg. Crowd Expertise	4	250
II	Heterogeneous Crowd	Var. Crowd Expertise	5	250

Figure 2.3: Crowd expertise distributions for Cases I and II that test how the expertise of evaluators within the crowd affect evaluation error for homogeneous and heterogeneous crowds, respectively. Three possible sample crowds are shown for both cases.

randomly; the location parameter γ captures evaluation losses intrinsic to the system, such as those stemming from the human-computer interaction.

Next, the variance $\mathbb{V}$ of the evaluator error is considered constant, capturing the notion that, while we hope the major variability in the evaluation error to be captured by Equation ((2.3)), other reasons exist to spread this error, represented by constant C in Equation ((2.4)).

$$\mathbb{V}\left[\sigma_{pd}^2\right] = \mathrm{C} \tag{2.4}$$

Following the requirements given by Eq. (2.3) and (2.4), we reparameterize the Inverse-Gamma of Eq. (2.2) to obtain Eq. (2.5) and (2.6).

$$\alpha_{pd} = \frac{1}{C\left(1 + e^{\theta(d_d - a_p) - \gamma}\right)^2} + 2 \tag{2.5}$$

$$\beta_{pd} = \left(\frac{1}{e^{\theta(d_d - a_p) - \gamma}}\right)\left(\frac{1}{Ce^{2\theta(d_d - a_p) - 2\gamma}} + 1\right) \tag{2.6}$$

The hierarchical random variables of the evaluator's evaluation expertise a_p and the design's evaluation difficulty d_d are both restricted to the range [0,1]. We let their distributions be truncated Gaussians with parameters μ_a, σ_a^2, μ_d, σ_d^2 set globally for all evaluators and designs as shown in Eq. (2.7) and (2.8).

$$a_p \sim \text{Truncated-Gaussian}\left(\mu_a, \sigma_a^2\right), \quad a_p \in [0,1] \tag{2.7}$$

$$d_d \sim \text{ Truncated-Gaussian}\left(\mu_d, \sigma_d^2\right), \quad d_d \in [0,1] \tag{2.8}$$

The probability densities over θ and γ are assumed as Gaussian with parameters μ_θ, σ_θ^2, μ_γ, σ_γ^2 as shown in Eq. (2.9) and (2.10).

$$\theta \sim \text{Gaussian}\left(\mu_\theta, \sigma_\theta^2\right) \tag{2.9}$$

$$\gamma \sim \text{ Gaussian}\left(\mu_\gamma, \sigma_\gamma^2\right) \tag{2.10}$$

Finally, by combining all random variables described in this section, we obtain the joint probability density function shown in Eq. (2.11).

$$\begin{aligned} &p\left(\mathbf{a}, \mathbf{d}, \Phi, \mathbf{R}, \theta, \gamma\right) = \\ &p(\theta)p(\gamma)\prod_{p=1}^{P} p(a_p)\prod_{d=1}^{D} p(r_{pd}|a_p, d_d, \theta, \gamma, \Phi_d)p(d_d)p(\Phi_d) \end{aligned} \tag{2.11}$$

Note that all hyperparameters are implicitly included.

Estimation and Inference of the Bayesian Network

The Bayesian network crowd consensus model is built upon the following random variables: Evaluators' expertises $\{a_p\}_{p=1}^{P}$, designs' difficulties $\{d_d\}_{d=1}^{D}$, true scores of designs $\{\Phi_d\}_{d=1}^{D}$, and parameters θ, γ, μ_a, σ_a^2, μ_d, σ_d^2. This section explains the settings for inferring the random variables and estimating the parameters using the observed evaluations of the evaluators $\mathbf{R} = \{r_{pd}\}_{p=1,\ldots,P;d=1,\ldots,D}$.

Two techniques are used in sequence. Maximum a posteriori estimation is performed using Powell's conjugate direction algorithm (Powell, 1964), a derivative-free optimization method, to get initial estimates of the parameters that maximize Equation ((2.11)). These point estimates are then used to initiate an adaptive Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm (Haario *et al.*, 2001; Gelfand & Smith, 1990; Patil *et al.*, 2010) that determines the estimates of all unknown parameters and infers posterior distributions of the random variables. The posterior sample size of the single-chained MCMC simulation is set to $2\text{x}10^5$, thinned by a factor of 2, with the first half discarded as burn-in.

2.4 Simulated Crowds Experiment

We now conduct an experiment to assess how the expertise distribution of the crowd affects the crowdsourced evaluation process using Monte Carlo simulations. There are two main goals of this experiment. First, we want to understand how crowds made up of different mixtures of high and low-expertise evaluators affect the crowd's combined scores of designs and the subsequent evaluation error from the true scores of the designs. Second, we want to understand the performance differences between the Bayesian network and by Averaging. Specifically of interest are the conditions under which the Bayesian network is able to find the subset of high-expertise evaluators within the crowd so that it can give greater weight to their responses.

There are two crowd expertise distribution cases we test, as shown in Figure 2.3. Case I is that of a homogeneous crowd, where all evaluators making up the crowd have similar expertise. The varied parameter in the homogeneous case is the average expertise of the crowd, thus testing the question: How well can a crowd perform if no individual evaluator can evaluate correctly or, alternatively, if every evaluator can evaluate correctly? Case II is that of a heterogeneous crowd, where the crowd is made up of a mixture of high and low-expertise evaluators. In this case, we fix the average expertise of the crowd to be low, so that most evaluators cannot evaluate designs correctly. Instead, the varied parameter in the heterogeneous case is the variance of the crowd's expertise distribution. This tests the question: How well can a crowd perform as a function of its proportion of high-expertise to low-expertise evaluators?

The procedure for these experiments is to use the Monte Carlo simulation environment to: (1) Generate a crowd made up of evaluators with expertise drawn from the tested expertise distribution (Case I or II), and a set of designs with true scores unknown to the crowd; (2) simulate the evaluation process by generating a rating between 1 and 5 that each evaluator within the crowd gives to each design; (3) combine the evaluator-level ratings into the crowd's combined score for each design using either the Bayesian network or by Averaging; and (4) calculate the evaluation error between the true scores of the designs and the combined scores from either the Bayesian network or by Averaging.

The simulation setup for these experiments consisted of 60 evaluators per crowd, as well as eight designs with scores drawn uniformly from the range [0,1] and evaluation difficulties $\{d_d\}$ set at 0.5 for all designs. The evaluation process for each evaluator is to rate all eight designs in the continuous interval [1,5] according to a deterministic equation given by the right hand side of Equation (3), with the location parameter γ set at 0 and the scale parameter θ set at 0.1. After the crowd's combined scores are obtained, either by the Bayesian network or by Averaging, the evaluation error

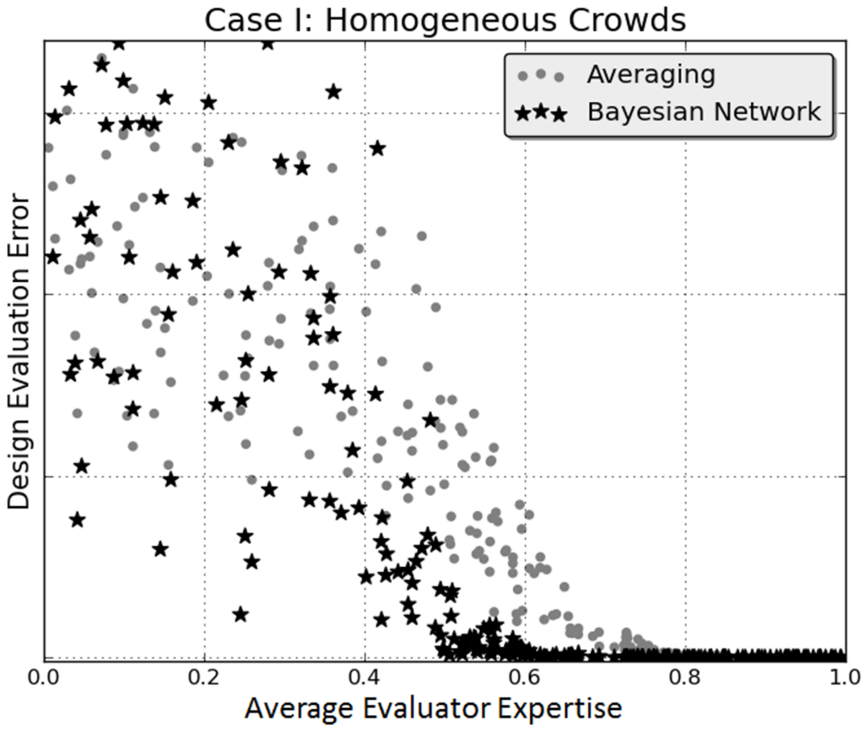


Figure 2.4: Case I: Design evaluation error from the Averaging and Bayesian network methods as a function of average evaluator expertise for homogeneous crowds. This plot shows that, when dealing with homogeneous crowds, aggregating the set of evaluations into the crowd's consensus score only sees marginal benefits from using the Bayesian network around 0.4 to 0.7 range of evaluator expertise.

between the combined scores $\hat{\Phi}_d$ and the true scores is calculated using the mean-squared error (MSE) metric as shown in Equation ((2.12)).

$$\text{MSE} = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\Phi}_d - \Phi_d\right)^2 \tag{2.12}$$

The results of Case I are shown in Figure 2.4. Each data point represents a distinct simulated crowd with average expertise given on the x-axis, and associated design evaluation error between the overall estimated score and the true scores on the y-axis. All crowds in Case I were generated using the same narrow crowd expertise variance $\sigma_a = 0.1$ to create homogeneous crowds. The results show that if the average evaluator expertise is relatively high, both Averaging and the Bayesian network

perform similarly with small design evaluation error. In contrast, when the average expertise is relatively low, neither Averaging nor the Bayesian network can estimate the true scores very well. Note that around an average evaluator expertise of 0.4 to 0.7, the Bayesian network performs marginally better than Averaging.

This observation agrees with intuition. A group of evaluators where "no one has the expertise" to evaluate a set of designs should not collectively have the expertise to evaluate a set of designs just by changing the relative weightings of evaluators and their individual evaluation responses upon combination when determining the crowd's combined score. Similarly, a group of evaluators where "everyone has the expertise" to evaluate a set of designs should perform well regardless of the relative weighting between evaluators. The key result for Case I is this: When the crowd has a homogeneous distribution of evaluator expertise, it does not significantly matter which weighting scheme one assigns between various evaluators and their evaluations; the Bayesian network and Averaging will perform similarly to each other.

The results of Case II are shown in Figure 2.4. Contrary to Case I, distinct crowds represented by each data point have on average the same expertise $\mu_a = 0.2$. Moving right along the x-axis designates crowds with increasingly higher proportions of high-expertise evaluators within the crowd. We observe that the Bayesian network performs much better than Averaging after a certain point on the x-axis; the point where a sufficient number of high-expertise evaluators is contained within the crowd. Under these conditions, the Bayesian network identifies the small group of experts from the less competent crowd and weighs their evaluation more than the rest, thus leading to combined scores much closer to the true scores of the designs. This observation is not present when the crowd does not have the sufficient number of high-expertise evaluators within the crowd. When this occurs, as is shown on the left side of the x-axis, the situation of "no one has the expertise" is recreated from Case I.

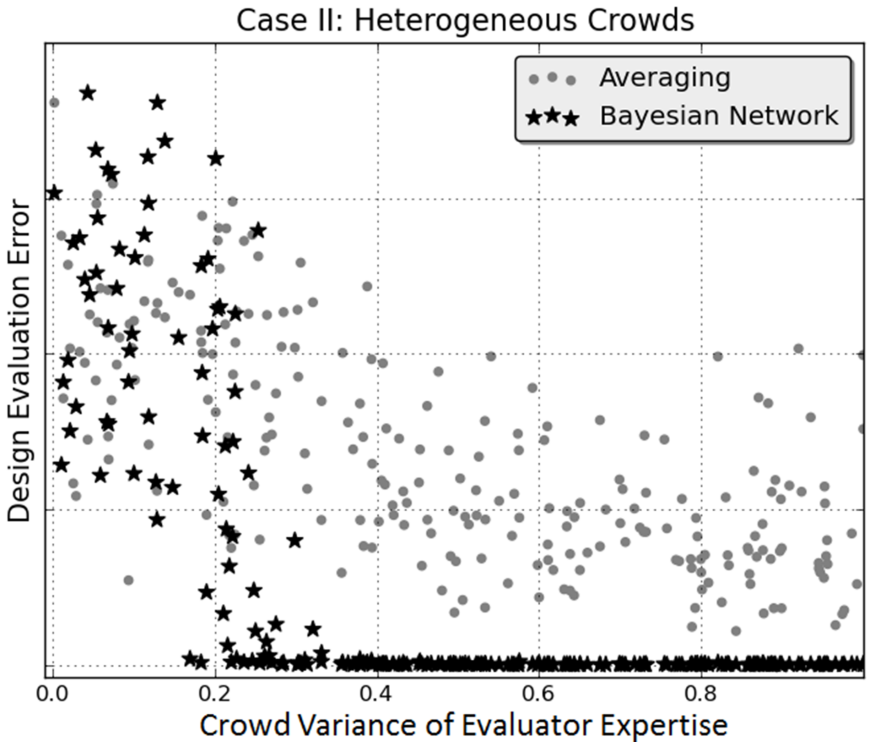


Figure 2.5: Case II: Design evaluation error over a set of designs for a mixed crowd with low average evaluation expertise. With increasing crowd variance of expertise there is an increasingly higher proportion of high-expertise evaluators present within the crowd. This leads to a point where the Bayesian network is able to identify the cluster of high-expertise evaluators, upon which evaluation error drops to zero.

In summary, we created simulated crowds to test the influence of crowd expertise on the crowdsourced evaluation process. Two cases were tested, representing homogeneous and heterogeneous expertise distributions. Under the modeling assumptions described in Section 2.3, we find that: (1) When the crowd is homogeneous, it does not matter what weighting scheme is used, as both Averaging and the Bayesian network give similar results; (2) when the crowd is heterogeneous, the Bayesian network is able to output the crowd's combined score much closer to the true scores under the condition that a sufficient number of "expert" evaluators exist within the crowd.

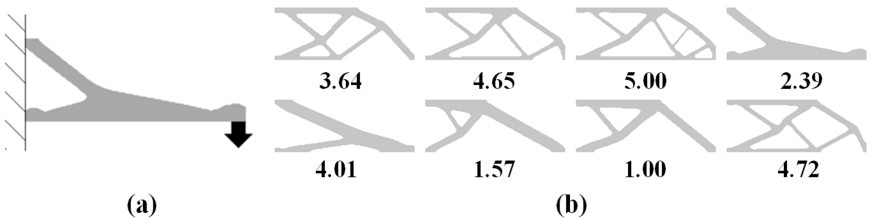


Figure 2.6: (a) Boundary conditions for bracket strength evaluation, (b) the set of all eight bracket designs

2.5 Human Crowd Experiment

In this section we test the performance of the Bayesian network crowd consensus model as compared with Averaging using an engineering design evaluation task and a real human crowd. The evaluation task was chosen to be a "simple" classic structural design problem for a load-bearing bracket (Papalambros & Shea, 2005), in which evaluators are asked to rate the capabilities of bracket designs to carry a vertical load as shown in Figure 2.4.

Participants

The human crowd consisted of 181 evaluators recruited using the crowdsourcing platform Amazon Mechanical Turk (Amazon, 2005). For the bracket designs, eight bracket topologies were generated using the same amount of raw material. The deformation induced by tensile stress upon vertical loading of each bracket was calculated in OptiStruct (Schramm *et al.*, 1999). The strength of a bracket was defined as the amount of deformation under a common load, and was subsequently scaled linearly between 1 and 5 as labeled in Figure 2.4. The scaled strength values were considered as the true scores, which were later used to calculate evaluation errors from the estimations from either the Bayesian network or Averaging methods.

Procedure

The evaluation process for each evaluator was as follows: The eight bracket designs were first presented all together to the evaluator, who was then asked to review these designs to get an overall idea of their strengths. After at least 20 seconds, the evaluator was allowed to continue to the next stage where the designs were presented sequentially and in random order. For each design, the evaluator was asked to evaluate its strength using a rating between 1 and 5, with 1 being "Very Weak" and 5 "Very Strong." To gather these data, a website with a database backend was set up that recorded when an evaluator gave an evaluation to a particular bracket design(University of Michigan - Optimal Design Laboratory, 2013).

Data analysis

A preprocessing step was carried out before the data were fed into either the Bayesian network or Averaging crowd consensus methods. Specifically, since some evaluators would give ratings all above 3 while some others tended to give ratings all around 3, all evaluations were linearly rescaled to a range of 1-5. It should be noted that while this mapping ensures that everyone gives '1's and '5's, it does not help to remove nonlinear biases in between an evaluator's most extreme evaluations. To calculate design evaluation error, the same mean-squared error metric was used as in the simulated crowd experiment and as given in Equation ((2.12)).

2.6 Results

The Bayesian network crowd consensus model did *worse* than Averaging when estimating the true scores of the bracket designs as shown in Table 2.1.

According to the simulation results, the Bayesian network can only do worse than Averaging if it is not able to find the experts in the crowd. This could happen under

	Design Evaluation Error (std.)
Averaging	1.001 (N/A)
Bayesian Network	1.728 (0.006)

Table 2.1: Mean-squared evaluation error and standard deviation from entire human crowd using Averaging and Bayesian network estimation.

either of the following two situations: (1) The modeling assumption made in Section 2.3 holds, namely, that low-expertise evaluators are less consistent (more random) in their evaluations, but there are just no high-expertise evaluators; (2) the modeling assumption is violated, in that there exist low-expertise evaluators consistently wrong in their evaluations. In this situation, the Bayesian network crowd consensus model would mistakenly identify evaluators as having high expertise due to their consistency and overweigh their incorrect evaluations.

Visualizing the crowd's expertise distribution

We now show that situation (2) above has occurred; namely, there are indeed "consistently wrong" evaluators that exist in the human crowd. To show this, we cluster the eight-dimensional human evaluation data to find clusters of similar evaluators, and then flatten these clustered data to two dimensions for visualization. This clustering finds groups of evaluators who give consistent evaluation, regardless of whether such evaluations are correct or incorrect. In other words, members of a cluster were consistent in their evaluations not necessarily to the right or wrong answer, but consistent to others in the cluster.

The clustering algorithm we used is density-based and uses the Euclidean distance metric to identify clusters of evaluators who gave similar evaluations (Ester *et al.*, 1996). This clustering method was chosen as it can account for varying clustering sizes, as well as not necessitating that every evaluator belong to a cluster. The flattening from eight dimensions to two dimensions was done using metric multidimensional scaling.

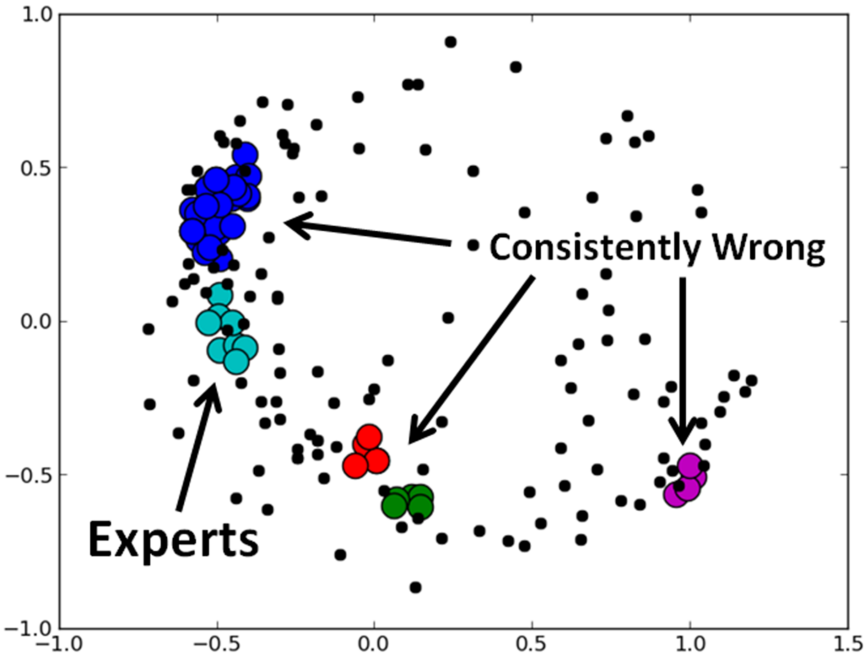


Figure 2.7: Clustering of evaluators based on how similar their evaluations are across all eight designs. Each black or colored point represents an individual evaluator, where colored points represent evaluators who were similar to at least 3 other evaluators, and black points represent evaluators who tended to evaluate more uniquely.

We see in Figure 2.6 that five clusters of similar evaluators were found, while Table 2.2 gives the evaluation error of each cluster. We find that the cyan cluster is made up of high-expertise "expert" evaluators, as evidenced by their evaluation error. In contrast, the other four clusters were "consistently wrong" in their evaluations.

This analysis suggests that finding expert evaluators through an open call is possible even for a task like structural design, in which expertise is sparsely distributed through the crowd. However, while the Bayesian network crowd consensus is a the-

Cluster Color	Design Evaluation Error
Blue	1.415
Cyan "Experts"	0.544
Red	1.652
Green	2.203
Magenta	6.031

Table 2.2: Mean-squared evaluation errors from the 5 clusters of similarly evaluators.

oretical way to identify these evaluators, its application in reality is limited by the fact that there exist other (more numerous) clusters of evaluators who are just as consistent yet wrong in their evaluations.

2.7 Additional experiments to assess what went wrong?

For completeness of the human crowd experiment, we conducted three follow-up experiments to capture the differences between the simulated crowd assumptions and results, and the human crowd results. The first follow-up experiment augments the human crowd data with simulated experts, in order to offset the "consistently wrong" evaluators with a larger cluster of experts. The second follow-up experiment tests the effect of removing the "consistently wrong" evaluators from the human crowd experiment. The third follow-up experiment remains entirely in simulation, and shows that the existence of enough "consistently wrong" evaluators will also cause the Bayesian network crowd consensus to fail to find experts in simulation as well, thus mimicking the results of the human experiment.

2.7.1 Human crowd augmented with simulated experts

We show in Figure 2.8 how the design evaluation error would be reduced if extra expert evaluators, i.e., evaluators with evaluations exactly the same as true scores, were collected in addition to the original 181 evaluators from the human experiment. Notice that the error should be reduced monotonically as the number of experts increases. However, the stochastic nature of the estimation process of a Bayesian network could cause sub-optimal estimations. Similar to the simulations in Figure 2.4, one can observe the phase-changing phenomenon in the change of the design evaluation error. This phase change represents when the Bayesian network is indeed able to find the experts in the crowd. Notice that although adding 10 additional experts does not make a majority of the crowd as expert, it is sufficient for the

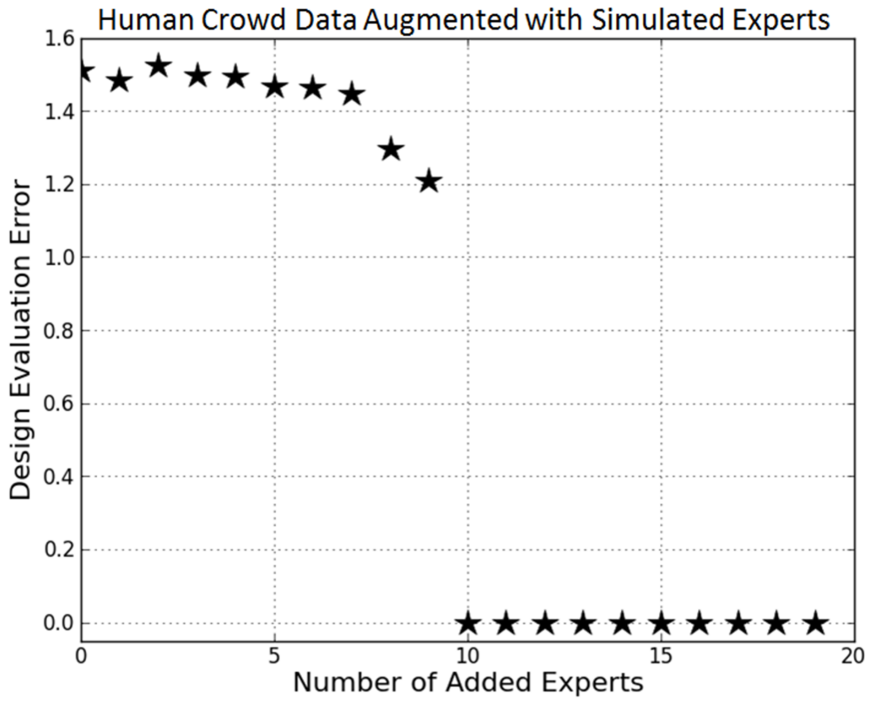


Figure 2.8: Design evaluation error with respect to additional experts.

Bayesian network crowd consensus model to locate the experts and subsequently overweigh their evaluations.

2.7.2 Human crowd with "consistently wrong" evaluators removed

We address how removing the "consistently wrong" evaluators affects the crowd's evaluation error, in which the "consistently wrong" evaluators are those found by clustering as shown in Figure 2.6. As reference, averaging the evaluations of the entire crowd results in a mean-squared error of 1.001 as given earlier in Table 1.

Removing the "consistently wrong" evaluators resulted in a *worse* evaluation error at 1.228 than averaging the entire crowd. This finding suggests that either the "consistently wrong" evaluators are not as wrong as the non-consistent non-experts (i.e., the humans that were not clustered as represented black dots in Figure 2.6), or that non-expert evaluation errors at the design level tend to cancel each other out.

It is found that indeed evaluation errors are being canceled at the design level.

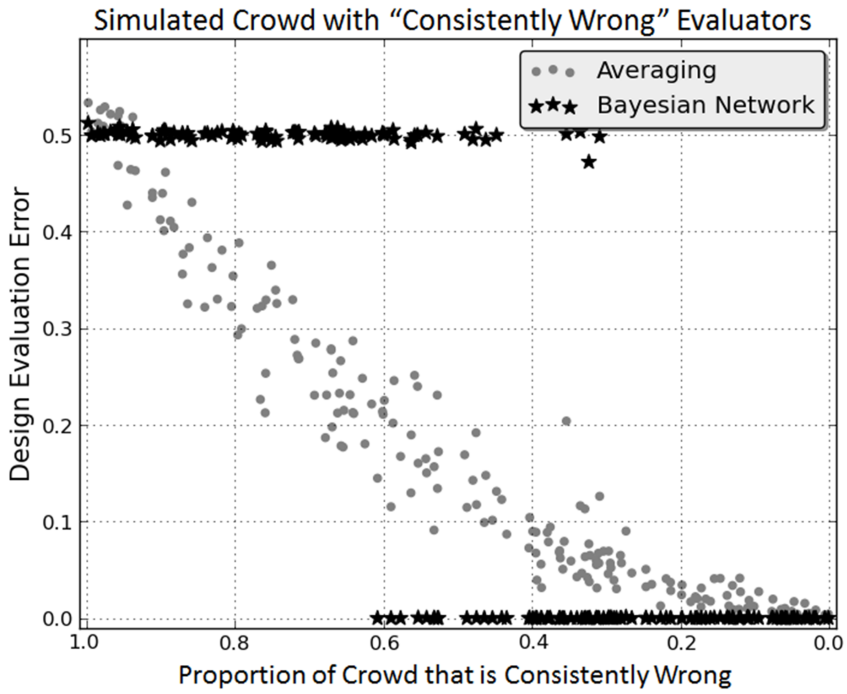


Figure 2.9: Design evaluation error with respect to the proportion of the expert group.

This is suggested by finding that the evaluation error of only the non-consistently wrong (black dots) is 1.339, while the evaluation error of both the consistent and non-consistently wrong (i.e., all but the experts) is 1.060. Note that the non-consistently wrong evaluators have an average evaluation error lower than that of any of the "consistently wrong" evaluators.

This analysis suggests that it is not sufficient, at least as far as this sample goes, to use the Bayesian network crowd consensus model to identify "consistently wrong" evaluators and simply omit them from the evaluation task. While their evaluations may obscure identification of the experts, they may be useful as they may be also canceling out errors from other evaluators.

2.7.3 Simulated crowd with "consistently wrong" evaluators

In this scenario, we tested a set of simulations in which the crowd contained two clusters of evaluators. One cluster, the "experts", can always evaluate correctly; the

other cluster is almost the same, except that evaluators in this cluster always rate one design consistently wrong by 0.5. We vary the crowd proportion of "consistently wrong" evaluators from 100% to 0% and calculate the corresponding evaluation errors as shown in Figure 2.7.2. While the error from Averaging changes linearly with respect to the proportion, that from the Bayesian network takes only two phases. The result mimics what we saw with the human experiment; the Bayesian network simply considers one of the clusters as the experts based on the cluster size and spread, regardless of whether that cluster is consistently correct or consistently wrong.

2.8 Summary

Crowdsourcing is a promising method to evaluate engineering design concepts that require human input, due to the possibility of leveraging evaluation expertise distributed over a large number of people. For engineering design tasks, a common characteristic of typical crowdsourced design evaluation processes is that the crowd is composed of a heterogeneous mixture of high and low-expertise evaluators. Simply averaging all evaluations from the crowd results in inaccurate crowd consensus scores for the set of designs, due to the large number of low-expertise evaluators. Consequently, a key challenge in such crowdsourced evaluation processes is to find the subset of expert evaluators in the crowd so that their evaluations may be given more weight.

In this chapter we developed and benchmarked a crowd consensus model in the form of a Bayesian network that aims to find the expert evaluators and subsequently give their evaluations more weight. The key modeling assumption for this crowd consensus model is that low-expertise evaluators tend to "guess," resulting in more random evaluations than expert evaluators.

We tested, using both simulated crowds and a human crowd, how the Bayesian network crowd consensus model performs compared to averaging all evaluations for

a "simple" engineering design evaluation task. We showed in simulation that when assumptions hold, the Bayesian network is able to find the experts in the crowd and outperform averaging. However, the results of the human crowd experiment show that we were *not* able to achieve a more accurate design evaluation using the Bayesian network crowd consensus model than just averaging all evaluations. It was found that there were numerous clusters of "consistently wrong" evaluators in the crowd, causing the Bayesian network to believe they were the experts, and consequently overweighting their (wrong) evaluations. These results suggest that crowd consensus models that *only* observe evaluations may not be suitable for crowdsourced evaluation tasks for engineering design, contrasting with many of the recent successes from the crowdsourcing literature.

Crowdsourced evaluation can fail for even a simple engineering design evaluation task due to the expertise distribution of the crowd; averaging already gives a low-accuracy estimate of design scores due to the large number of low-expertise evaluators, and crowd consensus models relying only on evaluations may not be able to find the experts in the crowd. Consequently, further research is needed into practical methods to find experts when they are only a small subset of the crowd as well as shrouded by numerous clusters of consistent yet incorrect evaluators.

Promising avenues in this direction may be in extending crowd consensus model to include relevant information to the engineering design evaluation task as has been done with item features (Raykar *et al.*, 2010), evaluator confidence (Prelec *et al.*, 2013), evaluator behavioral measures (Rzeszotarski & Kittur, 2011), and expertise assessed over longitudinal tasks (Budescu & Chen, 2014a). Another useful direction may be in analytic conditions for when experts in the crowd may be found (Della Penna & Reid, 2012; Waggoner & Chen, 2013; Davis-Stober *et al.*, 2014; Kruger *et al.*, 2014), possibly in the form of practical questions or tests to run before setting up an entire crowdsourced evaluation process. While this initial step displays potential challenges

for crowdsourced evaluation for even simple engineering design tasks, such extended crowd consensus models are likely to benefit a multitude of research communities.

CHAPTER III

Finding Experts in the Crowd using "Expertise Heuristics"

3.1 Context: How do we find the experts in the crowd?

In this chapter we again consider only objective design evaluation tasks, namely, tasks that possess a true score. For example, an objective task for the Chevrolet Volt would be to evaluate which hybrid powertrain architecture achieves the best fuel economy under emissions constraints (Bayrak *et al.*, 2013a). Contrast this with a subjective design evaluation task, such as asking a crowd of potential customers which aesthetic styling options should be offered for the same Chevrolet Volt design concept (Burnap *et al.*, 2015a).

The goal of a crowdsourced design evaluation is to aggregate the set of individual evaluations into a single combined evaluation, called the crowd consensus (Sheshadri & Lease, 2013b), that is as close to the (unknown) true score as possible. While methods of aggregating a number of evaluations into a single evaluation have been studied in many communities (e.g., scoring questions for aptitude tests (Embretson & Reise, 2013; Bachrach *et al.*, 2012b)), as well as continually utilized in a number of societal-level scenarios (e.g., voting for a democratically elected leader), there still exist open challenges when aggregating design evaluations.

As we saw in Chapter 2, a key issue that differentiates even "simple" engineering design tasks with other evaluation tasks is the relatively sparse distribution of expertise in the crowd (Peisl *et al.*, 2014; Burnap *et al.*, 2015b). Oftentimes complex engineering systems require specialized knowledge and experience for correct evaluation, namely, a high level of evaluation expertise, resulting in only a minority of experts in the crowd. This observation is supported by a number of business case studies, in which successful design evaluations are made by small groups within the crowd or even single evaluators (Chiu *et al.*, 2014; Diener & Piller, 2010; Peisl *et al.*, 2014); it is also supported in academic studies showing that "consistent, yet wrong" non-expert evaluators can overwhelm the relatively smaller set of "consistent, yet correct" expert evaluators resulting in a very incorrect crowd consensus (Burnap *et al.*, 2015b).

Accordingly, another stipulation in the present work is that we only consider difficult design evaluation tasks characterized by having a "minority of experts" in the overall crowd. This is in contrast to the more straightforward, yet relatively more academically researched, task of aggregating evaluations when the largest consistent group within the crowd is actually the experts themselves (Sheng *et al.*, 2008b; Sheshadri & Lease, 2013b). An example of such an evaluation task is image annotation, in which a large database of unannotated images is tagged according to objects in the image (Welinder *et al.*, 2010a). In this case, the task is simple, and "everyone is an expert;" performing a majority vote is sufficient to get a crowd consensus near the true scores of the designs (Sheng *et al.*, 2008b).

Given this overall narrowing of focus—aggregation of evaluations for objective engineering design evaluation tasks in which the experts are in the minority—a key challenge is to identify the experts in the crowd, such that non-experts and their evaluations may be filtered out (Peisl *et al.*, 2014). If engineering enterprises can successfully filter the minority of experts from the majority of non-experts, they can then

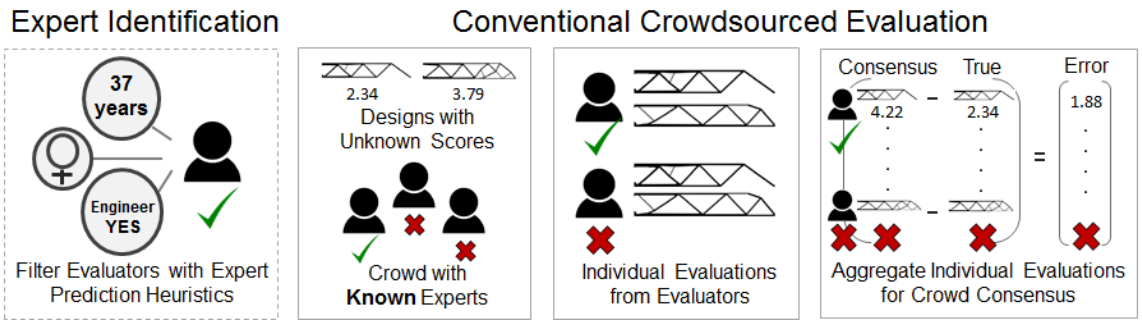


Figure 3.1: Overall flow of traditional crowdsourced evaluation process for an engineering enterprise. The enterprise starts with a set of designs with unknown true scores and a crowd with unknown experts. The crowd evaluates the designs and provides a score. Traditionally during the last step the crowd consensus is obtained by averaging all evaluations since expertise is unknown. The correct expert evaluations can be overshadowed by the incorrect non-expert evaluations. This research aims to add to the beginning of the process an additional expert identification stage to automatically identify a crowd with known experts, such that non-experts may be filtered out to improve the final crowd consensus evaluation.

aggregate the crowd consensus much closer to the objective true scores of the designs. With a filtered crowd of experts, complex engineering systems enterprises could incorporate crowdsourced evaluation methods into stage-gate reviews, improving their design processes and mitigating increased time and cost overruns.

Research Aim

Identifying and filtering expert from non-expert evaluators for a design stage-gate is not a new challenge, and has been studied under various names across a number of fields; for example: (1) Boeing's concurrent engineering processes involve creating cross-functional teams of experts from a number of relevant disciplines to evaluate design concepts at stage-gate reviews (Klein *et al.*, 2006); (2) IBM's InnovationJam used an expert crowd of 50 internally selected executives to evaluate innovative design concepts (Bjelland & Wood, 2008; Blohm *et al.*, 2013); and (3) Fiat's Mio used its internal engineering design team to evaluate crowdsourced design concepts (Celaschi *et al.*, 2011a). Common practice in these examples is that skilled managers identify

and filter expert evaluators to create the team or crowd.

The current research aims to create automation tools for expert identification in the crowd, to augment rather than replace the "manual" choices made by skilled managers. We give a high-level overview of this augmented crowdsourced design evaluation process in Figure 3.1. If expert prediction heuristics can be found, the automated filtering system can be placed at the beginning of a crowdsourced evaluation process as shown in Figure 3.1, thus keeping existing engineering stage-gate review workflows intact.

We conducted an experiment to automatically identify experts in the crowd. This experiment uses a standard "simple" problem in engineering design, topology optimization of a 2D bracket, where evaluators are asked to select the bracket topology that is strongest for given boundary conditions (Antonsson & Cagan, 2005; Papalambros & Chirehdast, 1990). Since we know the true score of each 2D bracket, we can correctly identify expert evaluators based on how consistently an evaluator correctly identifies the stronger bracket.

We provide some background in Section 3.2 and a description of the problem in Section 3.3. A pilot study is detailed in Section 3.4, followed by the experiment using four expertise prediction heuristics corresponding to four Research Questions (RQ) listed below. The pilot study calibrated the 2D bracket topology evaluation experiment and the experiment considered the following RQs:

1. Can we identify experts from evaluator demographics such as age, gender, education level, and performance self-critique? The hypothesis is that similar to a resume, one can identify experts from their personal information.

2. Can we identify experts from their evaluation reaction time behavior including average reaction time and variance in reaction time? The hypothesis here was that experts would solve the problems more quickly resulting in shorter average and smaller variance in reaction times.

3. Can we identify experts from their aptitude on seemingly-related mechanical reasoning tests? The hypothesis here is that experts would have greater aptitude and perform better on these tests.

4. Can we identify experts on an "easy known version" of the actual "difficult unknown" evaluation task? The hypothesis here is the same as in RQ 3, but with a different type of test.

The results show that we are unable to identify experts using traditional heuristics of demographics, reaction times, or seemingly-related mechanical reasoning aptitude tests, giving negative answers to RQ 1, 2, and 3. In contrast, a positive answer to RQ 4 was indicated.

3.2 Related Work

While the motivation for this work comes from business case studies from the open innovation and organizational management research communities detailing successes and failures implementing crowdsourcing processes (Huizingh, 2011; Peisl *et al.*, 2014), our research approach builds on studies regarding expertise within design teams from the engineering design community, as well as more general heuristics of expertise from the psychometrics community.

Expert Identification Heuristics

Expert identification heuristics refers to methods used to identify "experts" in a crowd, generally using some sort of testing procedure. In this work, we examine heuristics from studying expert and novice designers, general mechanical reasoning aptitude tests, and correlations with evaluator demographics.

Behavioral studies of designers, particularly expert designers and their differences from novices, have been a key focus of much design research in the last 30 years

(Dinar *et al.*, 2015). These studies often are involved with in-depth observation via ethnographic studies of the representation, thinking processes, and knowledge transfer of designers (Cross, 2004a). Much research in this direction is based on that of a single designer. Results have found that experts are "better" at representation, in which better is defined as level of detail and interconnectedness of current design knowledge (Björklund, 2013; Chai *et al.*, 2015), as well as previous knowledge as assessed through sketch recognition (Kavakli & Gero, 2001a) and of prior knowledge modeled using patent repositories (Fu *et al.*, 2013a) Another major point of divergence between experts and novices in design is experience, which lends itself to experts being more aware of pitfalls such as design fixation (Crilly, 2015; Moreno *et al.*, 2014).

For teams of designers, similar behavioral studies have been conducted. Yang studied single evaluator versus group consensus evaluation, and found that while single evaluators can make faster decisions, diverse group decisions often lead to better outcomes (Yang, 2010). The composition of the design team has shown that diversity leads to better designs (Lau *et al.*, 2012), as well as inter-team communication and "openness" (Telenko & Wood, n.d.). These studies have shown that diversity of demographics may be important for expertise. Design thinking has been found to be significantly different between experts and novices. Ho et al. found that experts tended to work backwards from the solution (Ho, 2001a). Expert designers are also found to not make "leaps" but more "hops" between analogies when traversing their internal design representation space (Ozkan & Dogan, 2013), perhaps due to the amount of short-term-memory required for variable design coupling (Flager *et al.*, 2014). When assessing ideas, expert designers may be more "breadth first" versus "depth first" searching of the design space (Cross, 2004a) and likewise tend to quickly prioritize design issues, while novices treat things equally (Ho, 2001a). Such factors may be integral in design thinking and conceptual speeds during design tasks, as is being pursued via the Applied Test of Design Skills (Shah *et al.*, 2013).

These various differences in design thinking and evaluation motivate us to capture reaction times for evaluators making evaluations, as a crude proxy for certain design thinking processes. Moreover, we ask evaluators to self-critique (also known as "meta-knowledge") their performance on the evaluation task. This approach has been shown to be useful to capture expert thinking processes on other non-engineering tasks as well, for example memorization of U.S. state capitols (Prelec & Seung, 2007; Prelec, 2004a).

Mechanical reasoning refers to innate or learned expertise in various mental tasks such as spatial manipulation and being able to correctly intuit the dynamics of a physical system. Many standardized tests of mechanical reasoning aptitude have been used to asses both students and practitioners; in particular, we are guided from previous results correlating mechanical reasoning aptitude and undergraduate student grades in engineering design courses (Field, 2007a) and physics courses(Kozhevnikov *et al.*, 2007).

With respect to spatial abilities, we used the standardized Mental Cutting Test (1939) and the Purdue Visualization of Rotations Test (Bodner & Guay, 1997; Vandenberg & Kuse, 1978). For our dynamics tests, we follow results provided by McKenna and Agogino (McKenna & Agogino, 2004) and Hegarty (Hegarty, 2004) showing the use of rope and pulley dynamics, Kozhevnikov et al. for the use of a kicked ball (Kozhevnikov *et al.*, 2007), and Hegarty for 2D and 3D intermeshed gear rotations (Hegarty, 2004).

3.3 Problem Formulation: Models of Expertise Prediction

In this section, we describe the "simple" engineering design evaluation task, as well as notions of expertise and difficulty necessary to carry out experiments and present results of our tested crowdsourced design evaluations.

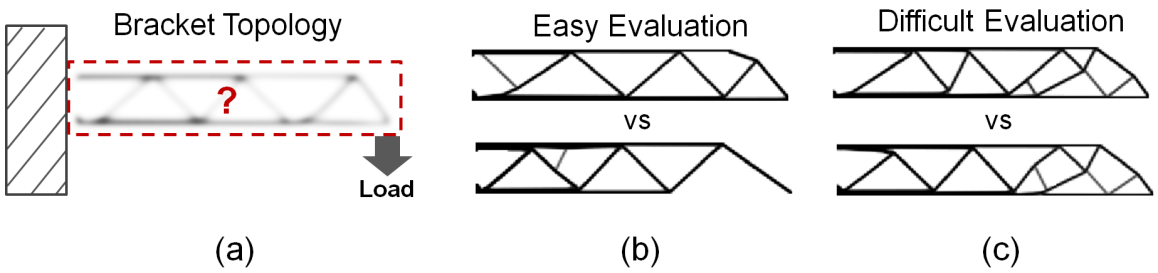


Figure 3.2: Overall flow of traditional crowdsourced evaluation process for an engineering enterprise. The enterprise starts with a set of designs with unknown true scores and a crowd with unknown experts. The crowd evaluates the designs and provides a score. Traditionally during the last step the crowd consensus is obtained by averaging all evaluations since expertise is unknown. The correct expert evaluations can be overshadowed by the incorrect non-expert evaluations. This research aims to add to the beginning of the process an additional expert identification stage to automatically identify a crowd with known experts, such that non-experts may be filtered out to improve the final crowd consensus evaluation.

Case Study: Bracket Topology Design Optimization

We used the bracket design evaluation task for all studies in this chapter(Antonsson & Cagan, 2005; Papalambros & Chirehdast, 1990); namely, given rectangular boundary conditions and a constant amount of material, which topology is able to hold the most weight at its tip? As shown in Figure 3.3, boundary conditions are setup to secure a bracket against a fixed support with the goal of optimizing a bracket topology for supporting a tip load using a constant amount of material. This toy problem has a very large (2ˆC, where C is the number of finite-elements) set of possible design concepts, characterized by very nonlinear regions of bracket strength when moving around the design space, i.e., small perturbations in bracket topology often result in large changes in bracket tip loading strength.

We provide two example bracket design evaluation tasks in Figure 3.3. For example, given the same amount of material, the top design is stronger than the bottom design in Figure 3.3(b); while this comparison is not as straightforward for the bracket pair shown in Figure 3.3(c).

Evaluation Expertise

We assume there is a single metric of expertise for the bracket design evaluation task. However, unlike a large body of previous work (e.g., (Bachrach *et al.*, 2012b; Burnap *et al.*, 2015b; Welinder *et al.*, 2010a)), we do not aim to explicitly estimate this expertise as a value prescribed by some assumed metric. Instead, we define expertise by proxy through what we are actually interested in—evaluation accuracy. Similar to how we pretend not to know the true score of bracket designs during evaluation, we pretend not to know who is an expert. Accordingly, we use evaluation accuracy, or how many correct binary evaluations an evaluator makes divided by the number of total binary evaluations, as the value of evaluation expertise.

Evaluation Difficulty

As stated in Section 3.1, one contribution of this work is the use of a controllable evaluation difficulty during the design of experiments, thus helping to better measure evaluator expertise by removing evaluation difficulty from the equation. Previous studies have measured evaluator expertise while assuming all designs are equally difficult to evaluate (Dawid & Skene, 1979), or by jointly inferring design difficulty along with evaluator expertise (Bachrach *et al.*, 2012b), thus posing potential statistical unidentifiability problems (Lakshminarayanan & Teh, 2013).

In this work, we algorithmically generate bracket designs and bin them into a histogram according to their true score as measured by bracket loading strength. These brackets are then presented in pairs to evaluators, followed by binary choice evaluation corresponding to which bracket the evaluator believes is stronger. We control evaluation difficulty by selecting brackets at varying "bin differences." A pair of brackets separated by many bins, or equivalently with very large differences in bracket strength, is easier to evaluate than a pair of brackets from adjacent bins. We give an example of two bracket pairs in Figure 3.3(b) and 3.3(c) showing how

evaluation difficulty may be parameterized by bracket strength bin difference.

However, mapping evaluation difficulty to bracket bin difference must be calibrated. While we intuitively know that small bin differences result in a more difficult evaluation task, and equivalently large bin differences result in an easier evaluation task, for the purposes of the experiment, we must know what this relationship looks like quantitatively. Consequently, we conducted a pilot study to discover this relationship. Using the results of this pilot study, a new crowd of evaluators was gathered for the now calibrated Experiment, as will be detailed in Section 3.4.

3.4 Hypotheses and Experiments

The experiment consisted of an initial pilot study to calibrate design difficulty, followed by an experiment consisting of 4 studies to test each of the research questions posed in Section 3.1.

3.4.1 Pilot Study: Calibrating Design Difficulty

The goal in the pilot study is to calibrate the 2D bracket topology evaluation task. Since we seek a controlled method of varying design difficulty, we generate bracket topologies for various boundary conditions and calculate true scores for load performance; thus, we can control how the magnitude and distribution of bracket design difficulty is presented to evaluators. We generated thousands of bracket designs, randomly assigned a crowd of evaluators to pairs of brackets, and let their evaluations determine what is difficult to evaluate and what is easy to evaluate on average.

3.4.1.1 Evaluators

A total of 272 evaluators were sourced for the Pilot Study using the crowdsourcing platform Amazon Mechanical Turk, in which evaluators were given a monetary incentive for completing a predefined number of 2D bracket topology evaluation

tasks. These evaluators were given randomly selected bracket bins for evaluation. Accordingly, most bracket pairs clustered around probabilistically likely draws—akin to rolling the sum of two die and rolling a 7 than either a 2 or a 12. Accordingly, for the Pilot Study, we required evaluators see at least 6 unique bracket bin differences, with a median bracket evaluation time between 3 and 10 seconds. This resulted in a filtered set of 34 evaluators. This means that most of the original participants did not see a sufficient number of unique bracket difficulties (e.g., bin difference of 3 and 4).

Note that this set of evaluators was only used for the Pilot Study to calibrate bracket evaluation difficulty. They were not used for the Experiment as detailed in Section 5. Moreover, bracket bin differences were forced to be uniform during the subsequent Experiment, thus retaining the majority of evaluators in contrast with this Pilot Study.

3.4.1.2 Designs

A total of 4,829 designs were generated according to 2D boundary conditions on the left side as well as the lower-right tip as shown in Figure 2(a). These brackets were generated using an open-source topology optimization software (Andreassen *et al.*, 2011) with an element-binning of 250 units wide and 40 units tall, 20% of the bounding area containing mass and 80% no mass, Young's modulus of 1.0 N/mm2, and Poisson ratio of 0.3. To obtain variability in the bracket topologies, an additional random boundary condition on the interior of the bracket volume was added during the generation of each design. This random component had the same magnitude, but a different angle than the tip boundary condition.

The true score of each of these designs was assessed via finite-element analysis (Andreassen *et al.*, 2011). True scores were defined as the deflection at the lower-right tip of the bracket with a range between 2,980 and 1,238,533 and with a highly-

skewed distribution. In order to have roughly similar numbers of brackets of various true scores for later bracket evaluation, these bracket designs were subsampled to obtain a uniform distribution into 10 bins. This subsampling process proceeded by sampling 100 brackets from the first 10 bins, resulting in a total of 1,000 brackets filtered from the original 4,829.

3.4.1.3 Procedure

A web-based interactive survey was created to collect evaluations. Evaluators first visited a home page, where the experiment background and experiment instructions were provided. Evaluators were told that they would be given 18 bracket design evaluations, followed by 8 mechanical reasoning tests, followed by a demographic survey.

After clicking the mouse to proceed to the evaluations, a randomly selected pair of brackets was presented to the evaluator for binary comparison. For the Pilot Study, the random selection process involved randomly selecting a bracket from a random bin between Bin 1 and Bin 10. Bins were uniquely selected such that each bracket pair consisted of two brackets from separate bins. The bracket difference was recorded according to the absolute distance between the bins.

3.4.1.4 Pilot Study Results

The goal in the Pilot Study was to calibrate the evaluation difficulty for various bracket design pairs. As shown in Figure 3.4.1.3, we found that brackets with a "bin difference" of 1-5 result in low average crowd accuracies. Importantly, these pilot results show that the relationship between average accuracy and bin difference is relatively linear, suggesting our use of bin difference as a proxy for evaluation difficulty is useful during the design of the Experiments for RQ 1, 2, 3, and 4. These results are echoed by plotting the distribution of individual evaluator accuracies for

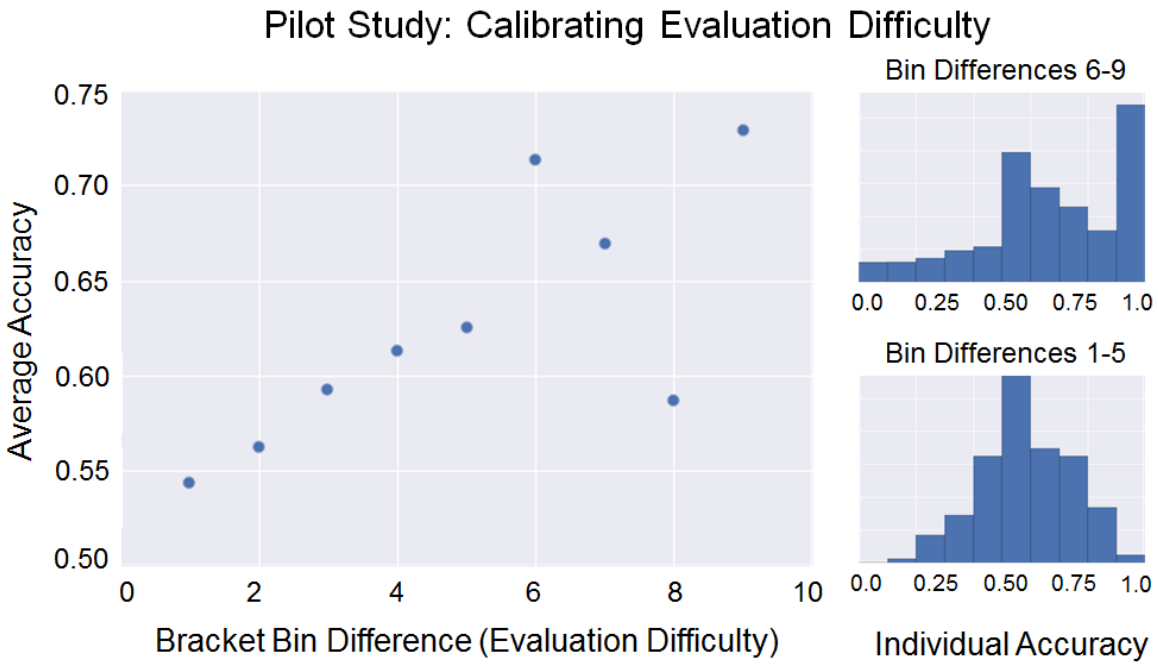


Figure 3.3: Pilot study results used to calibrate bracket difficulty for design evaluations in Studies 1, 2, 3, and 4. By taking the "bin difference" from the 10 true strength bins, we show brackets that are relatively similar in true strength are much more difficult to evaluate, as evidenced by average crowd evaluation being close to a random guess (0.50). In contrast, large bracket bin difference result in higher average accuracy. We thus chose to calibrate our experiment by splitting bracket bin differences from 1-5 and 6-9, as we can see individual accuracy is symmetrically distributed for the bin differences of 1-5.

bin difference 1-5 and 6-9. As is shown in Figure 3.4.1.3, evaluation tasks with a bin difference between 6-9 show many evaluators get significantly above random guesses of 50% accuracy, while bin difference of 1-5 show evaluations symmetrically distributed around an accuracy of 50%.

3.4.2 Demographics, Task Behavior, Mechanical Reasoning, and Using an Easy Task to Predict Expertise on the Actual Hard Task

Our goal in the Experiment was to ask four research questions, corresponding to four expertise prediction heuristics, in an attempt to filter expert evaluators from non-expert evaluators. The four research questions assess expert identification heuristics

using demographics, reaction times, mechanical reasoning aptitude, or an easy known version of the hard unknown evaluation task.

3.4.2.1 Evaluators

A crowd of 398 evaluators were sourced for the Experiment. These evaluators were different from those evaluators in the Pilot Study, thus constituting a new crowd. Further this new crowd was used for Research Questions 1, 2, 3, and 4. The 398 evaluators were filtered according to the same criteria used in the Pilot Study. In particular, all participants that did not fully complete the demographics survey, or those who had a median time bracket evaluation time less than 2 seconds or greater than 10 seconds, were removed. After this filtering process, a total of 334 evaluators constituted the crowd, which is a significantly greater retention ratio than were retained in the Pilot Study.

3.4.2.2 Designs

The same set of 1,000 designs generated and subsampled for the Pilot Study, as detailed in Section 4, was used in the Experiment.

3.4.2.3 Research Question 1 Variables: Demographics

The hypothesis tested by Research Question 1 is that demographics can be used to identify expert evaluators for the 2D bracket evaluation task. A total of 5 demographic variables were used in this research question: self-critique, age, gender, education level, and whether the evaluator was an engineer. Note that all of these variables are self-reported, and may not accurately reflect the true status of the evaluator. These demographic variables were tested for statistically significant correlation with evaluation expertise, and thus subsequently used for identifying experts in the crowd to improve the crowdsourced evaluation process.

1. The self-critique demographic variable refers to how well an evaluator felt she/he did at evaluating bracket designs. This variable was selected as previous research has shown self-critique may be used for identifying expertise for tasks such as remembering U.S. state capitals (Prelec & Seung, 2007). This variable had an integer range from 1 to 5.

2. Age refers to the number of years that the evaluator has been alive. The age demographics variable had 5 categorical options consisting of age ranges that were then converted to sequentially increasing integers for analysis.

3. Gender refers to the self-identified gender of the evaluator. The gender demographic variable was binary, and was converted to 0 or 1. These variables included a "Prefer not to say" option, in which case the evaluator was filtered out.

4. Education level refers to the highest level of education an evaluator has achieved, including currently enrolled students. This variable had 4 options: Some high school, high school graduate, some college, college graduate, some graduate school, and graduate school graduate. This demographic variable was converted to sequentially increasing integers for analysis.

5. The engineer demographic variable refers to whether an evaluator was currently working in an engineering field or was in school to be an engineer. This variable was binary, and was converted to 0 or 1.

3.4.2.4 Research Question 2 Variables: Reaction Time

The hypothesis tested by Research Question 2 is that evaluator reaction times can be used to identify expert evaluators for the 2D bracket evaluation task. The reaction time, defined as the time from the when the bracket pair are presented to the time the evaluation is submitted, was recorded. A total of 3 reaction time variables were used

in this research question: mean evaluation time, median evaluation time, and variance in evaluation time. These response time variables were tested with a goal of significant correlation with evaluation expertise, and thus subsequent use for identifying experts in the crowd to improve the crowdsourced evaluation process.

3.4.2.5 Research Question 3 Variables: Mechanical Reasoning

The hypothesis tested by Research Question 3 is that mechanical reasoning aptitude can be used to identify expert evaluators for the 2D bracket evaluation task. In this research question, we attempt to identify experts by using standardized mechanical reasoning aptitude tests from the psychometrics and mechanical reasoning testing communities. As detailed in Section 2, a number of studies have correlated mechanical reasoning aptitude with engineering student success as defined by grades (Bodner & Guay, 1997; Hegarty, 2004; Kozhevnikov *et al.*, 2007).

We choose five categories of standardized mechanical reasoning tests: block cutting, spatial rotation, gear rotation, dynamics, and a combination of all. Each of the first 4 categories consisted of two multiple choice questions, leading to a total of 8 mechanical reasoning questions given to each evaluator. The 5th category was a composite score of the responses from the 8 questions. We give a description of all 5 categories below, with references to their original standardized tests.

1. The Mental Cutting Test (Field, 2007b) presented evaluators with a 3D block that was bisected with a 2D plane. Evaluators were then asked what the corresponding cross section of the bisected block would look like from a viewpoint orthogonal to the cutting plane. Each block cutting test had 5 possible multiple choice options, with only one correct answer.

2. The Purdue Visualization of Rotation test (Vandenberg & Kuse, 1978; Bodner & Guay, 1997) presents evaluators with two 3D blocks, hereafter referred to as Block A and Block B. Block A was rotated along multiple axis, with its new

orientation A' shown next to the original orientation A. Evaluators were asked to rotate Block B with the same rotation as Block A. Overall, these tests were of the form, "A is to A' as B is to ___?" Five possible rotations of the Block B were provided, with only one correct answer.

3. The Gear Rotation Test showed evaluators either a 2D or 3D set of numerous intermeshed gears. Evaluators were asked which direction the last gear would turn after providing rotation to the first gear in the sequence. Note that inherently this mechanical reasoning variable had binary options, corresponding to either clockwise or counter-clockwise rotation.

4. The Dynamics mechanical reasoning variable consisted of two tests. The first test asked evaluators which possible flight path would be taken for a kicked ball (Kozhevnikov *et al.*, 2007). This test had 5 possible options. The second test asked evaluators which direction a pulley would spin (Hegarty, 2004), given that it was intertwined with a number of other pulleys and masses. Similar to the gear rotation test, this test only had binary options corresponding to either clockwise or counter-clockwise rotation.

5. The 'all mechanical reasoning' variable gave a uniform combination of all four previous mechanical reasoning variables. In other words, no new questions were asked, just an averaging of previous mechanical reasoning test scores.

3.4.2.6 Research Question 4 Variable: Easy Known Task

The hypothesis tested by Research Question 4 is that performance on an easy known evaluation task can be used to identify expert evaluators for the 2D bracket evaluation task. The easy known version of the evaluation task refers to a scaled down version of the of the hard unknown evaluation task that the engineering enterprise is actually after. Figure 3.3(b) given an example of an easy known version as compared

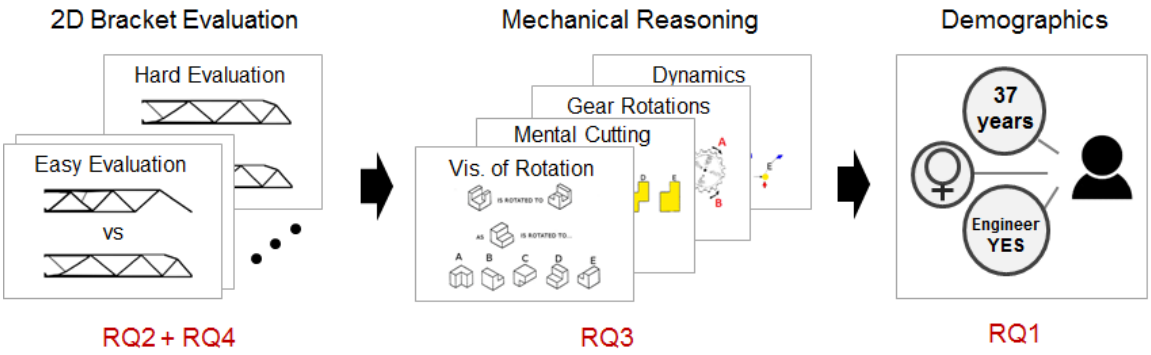


Figure 3.4: Diagram of the experiment procedure in order shown to evaluators. Three stages comprised of bracket evaluation and expertise assessment, mechanical reasoning tests, and demographic questionnaire were presented in sequential order. Within each stage, evaluation pairs or tests were presented randomly. The corresponding research question for each stage is highlighted in red.

to an example of the hard unknown evaluation task given in Figure 2(c). In other words, we selected only brackets with bracket difficulty with bin difference greater than 6 as shown in Figure 3.4.1.3. The evaluator accuracy on the easy known versions of the bracket evaluation task acted as a single variable for predicting evaluator expertise.

Experimental Procedure

We use the same crowd for all four research questions in the Experiment, and give a high-level overview of the procedure in Figure 3.4.2.6. A web-based interactive survey was created to collect evaluations. Evaluators first visited a home page, where the experiment background and experiment instructions were given. Evaluators were told that they would be given 18 bracket design evaluations, followed by 8 mechanical reasoning tests, followed by a demographic survey.

After clicking the mouse to proceed to the evaluations, a randomly selected pair of brackets was presented to the evaluator for binary comparison. For the Pilot Study, two brackets selected randomly from the 10 bins were presented. In the Experiment, however, each bracket pair consisted of two brackets from separate bins. Within each

bracket pair, one bracket design was from either Bin 1 or Bin 10. The evaluation task difficulty was recorded as the absolute distance between the bins. Always selecting one of bracket from either bin 1 or bin 10 ensured that the distribution of evaluation difficulties was relatively uniform for each evaluator.

Evaluators were asked to choose the stronger of the two bracket design topologies for holding a vertical tip load by clicking their mouse on one of the two presented bracket designs. Evaluators were allowed to change the selected bracket before choosing to submit a given pair, but were not allowed to go back and change previously selected brackets. Evaluation time, defined as the time from the when the bracket pair are presented to the time the evaluation is submitted was recorded.

After each evaluator completed 18 randomly generated bracket design evaluations, they then proceeded to the mechanical reasoning portion of the experimental procedure. Evaluators were then presented mechanical reasoning test questions one at a time, but in a random test question order (i.e., permutation of the integers 1-8). Mechanical reasoning test questions were either binary choice or multiple choice. Finally, evaluators proceeded to a demographic survey consisting of 6 demographic questions. After completing these demographic questions, evaluators were given a code to redeem a cash incentive.

3.5 Results

Research Question 1 Results: Demographics

We were not able to find any demographic variables that significantly correlated with evaluation expertise. This is shown visually in Figure 3.5, in which correlation coefficients between all demographic variables and expertise are given.

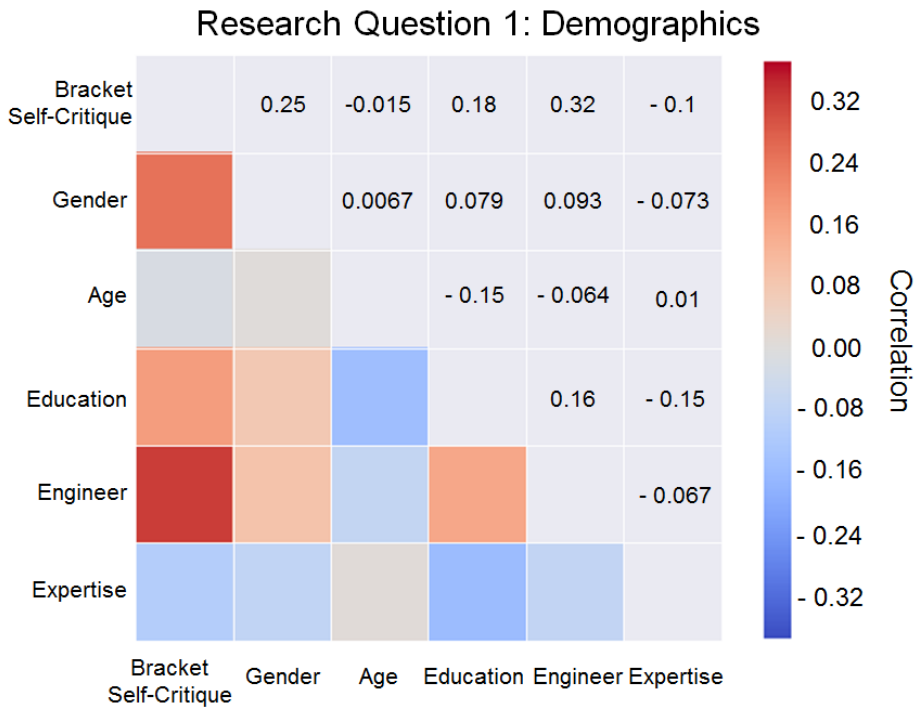


Figure 3.5: Experimental results of correlating demographics with expertise for Research Question 1. As can be seen by the bottom row of the correlation plot, expertise does not correlate significantly with any of the tested demographics.

Research Question 2 Results: Reaction Times

We were not able to find any reaction time variables that significantly correlated with evaluation expertise. This is shown visually in Figure 3.5, in which correlation coefficients between all reaction time variables and expertise are given.

Research Question 3 Results: Mechanical Reasoning

We were not able to find any mechanical reasoning categories that significantly correlated with evaluation expertise. This includes the additional fifth mechanical reasoning variable corresponding to the combined accuracy of all four individual mechanical reasoning categories. This is shown visually in Figure 3.5, in which correlation coefficients between all mechanical reasoning categories and expertise are given.

Figure 3.6: Experimental results of correlating evaluation reaction time with expertise for Research Question 2. As can be seen by the bottom row of the correlation plot, expertise does not correlate significantly with any of the tested reaction time variables

Research Question 4 Results: Easy Known Evaluation Task

We were able to find significant correlation between how well evaluators performed on the easy known version and evaluation expertise. Figure 3.5 shows the raw scatter plot, with the same scatter plot with additional jitter just for visualization purposes due to many data points being stacked upon each other. Additional visualization of the relationship between easy known version accuracy and expertise is given in Figure 3.5, in which the mean and variance are plotted, followed by plotting ordinary linear regression to convey the general trend between using an easy known version of the evaluation task and evaluation expertise on the actual unknown evaluation task. Note that the slope and intercept governing this trend is likely limited to the particular evaluation task of binary choice on bracket topology designs.

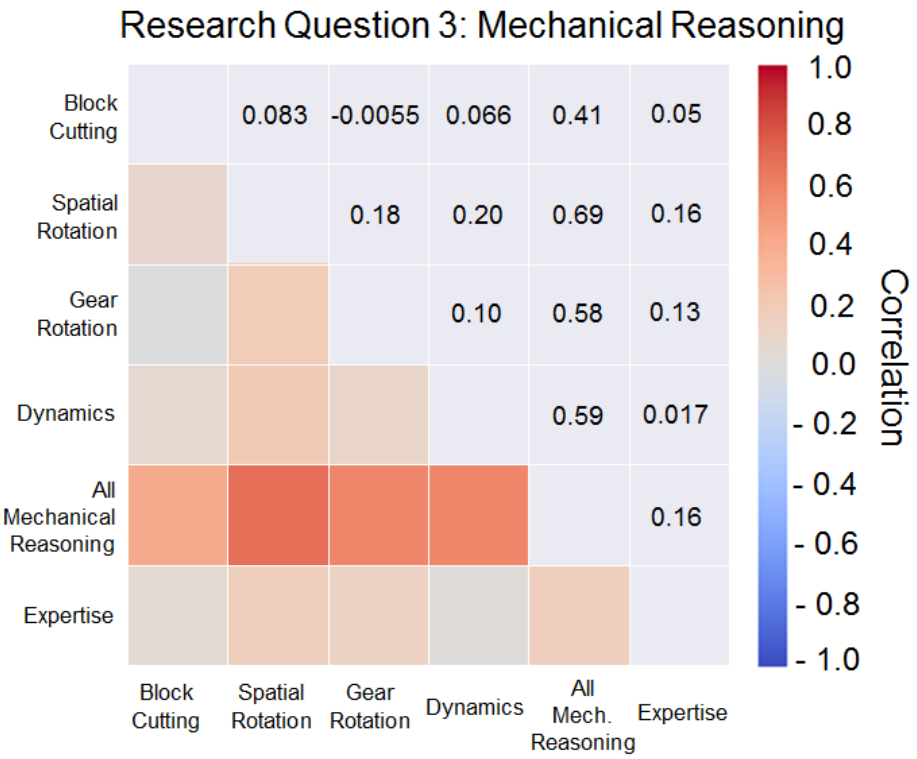


Figure 3.7: Experimental results of correlating mechanical reasoning aptitude with expertise for Research Question 3. As can be seen by the bottom row of the correlation plot, expertise does not correlate significantly with any of the tested mechanical reasoning categories.

3.5.1 Practical Usage: Identifying Experts to Improve Crowd Aggregation

As stated in Section **??**, the research aim of this work is to provide complex engineering systems enterprises a practical method for improving the crowd consensus evaluation by adding a preliminary expert filtering step as shown in Figure 3.1. Accordingly, we use the only practical method we found in our four Research Questions—the use of an easy known version of the actual unknown evaluation task—to show the degree of improvement on the bracket evaluation task used in this research.

In particular, we used the same crowd of 334 evaluators from Research Questions 1-4, and filtered out those who had an evaluation accuracy of less than 75% on the easy known evaluation task. This expert filtering left a subset of 134 experts. The

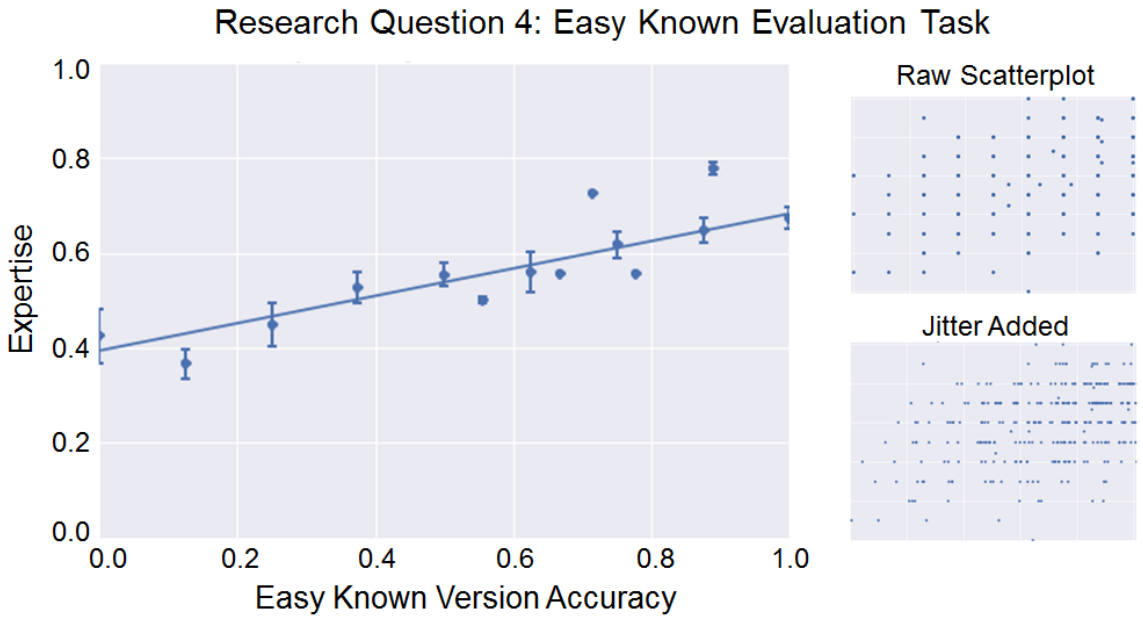


Figure 3.8: Experimental results of assessing accuracy on an "easy known evaluation task" with for Research Question 4. As can be seen in the linear regression plot, expertise accuracy on an easy known version of the evaluation task exhibits a positive trend with evaluation expertise. Also plotted are the raw scatter plot and "jittered" scatter plot since many data points lay on top of each other.

table below shows the average evaluation accuracy of the expert filtered crowd as compared with the original unknown crowd.

This accuracy improvement is promising, particular given that the crowd used in this research was recruited from a crowdsourcing website. More significant improvements in evaluation accuracy are possible given an initial crowd that is more likely to have expertise. In particular, "internal" crowdsourcing within a large engineering

Crowdsourcing Method	**Crowd Consensus Evaluation Accuracy**
Conventional Crowdsourced Evaluation	59.77%
Conventional Crowdsourced Evaluation + Expert Filtering	66.16%

Table 3.1: Practical usage of filtering experts to obtain improved crowd aggregation evaluation accuracy.

company is more likely to have relevant expertise (Erickson *et al.*, 2012). Business case studies have suggested that these internal crowds may not only gather expertise from non-traditional stakeholders, but offer a method of bypassing bureaucratic barriers (Erickson, 2013). Readers are referred to survey papers (Chiu *et al.*, 2014; Peisl *et al.*, 2014) and recent thesis (Erickson, 2013) for more discussion of enterprise benefits of crowdsourcing, particularly as used internally within an enterprise.

Furthermore, this work only focused on finding which variables correctly identify experts. These variables, in particular accuracy on an "easy known task," may be input to more advanced statistical models to further improve the crowd consensus. While out of the scope of this current work, readers are referred to recent models that capture evaluator expertise as a function of these variables (Budescu & Chen, 2014b; Miller *et al.*, 2014; Raykar *et al.*, 2009).

3.6 Summary

Crowdsourced evaluation is method of aggregating human input from evaluators outside the conventional design process, thus leveraging additional expertise during design stage-gates that may help mitigate increasing cost and time overruns characteristic of engineering systems enterprises. While successes of crowdsourced evaluation have been documented by a number of business case studies, a key challenge for engineering systems enterprises is to identify and filter expert evaluators from non-expert evaluators, so that the combined crowd consensus evaluation is closer to the true scores of designs concepts.

We conducted an experiment to identify expert evaluators in the crowd using four expertise prediction heuristics found in the engineering design and psychometrics communities. In particular, these four heuristic, corresponding to four research questions, assessed the correlation between evaluator demographics, reaction times, mechanical reasoning aptitude, and accuracy on an "easy known version" of the ac-

tual unknown evaluation task. Using an online web survey, our experiment used a real crowd of 334 evaluators to evaluate the loading strength of various 2D bracket topologies.

The results showed that identifying experts is not correlated with traditional expertise prediction heuristics, as we were not able to find correlation between evaluator expertise and demographics, reaction times, or mechanical reasoning aptitude. Instead, we found that evaluator accuracy on the "easy known version" of the 2D bracket evaluation task was able to identify experts on the actual "hard unknown version" of interest. We showed that automatically filtering experts increases the combined crowd consensus evaluation over conventional crowdsourced evaluation. This automatic expert identification and filtering stage offers an additional tool to management, alongside manual selection of expert evaluation teams, to help incorporate expertise into early stage-gates of the engineering systems design process.

CHAPTER IV

Do we need to Filter Experts for Subjective Preferences?

4.1 Context: Balancing Design Freedom and Brand Recognition

When developing the next generation of an existing vehicle model, an automotive designer must balance tradeoffs between two competing customer considerations. One consideration is the customer's desire for novelty, as the appeal of the current model tends to fade with time (Martindale, 1990; Coates, 2003). The extent the designer is able to reach toward increasingly novel designs, in other words by deviating from past designs, defines the amount of *design freedom* available to the design team. Another consideration is the customer's desire for consistency with past designs, which can play an important role in *brand recognition*. Much as there is family resemblance among members of a family, the designer seeks to maintain a recognizable brand character among all the brand's members. Any deviation from the past may reduce the new design's association with the brand, as well as how it conveys design attributes known to be important to the customers (e.g., luxuriousness) (Aaker & Keller, 1990).

At the enterprise level, both design freedom and brand recognition are known to contribute significantly to market competitiveness (Bloch, 1995; Person *et al.*, 2007;

Yin Wong & Merrilees, 2008). On the academic side, studies have shown that vehicle manufacturers that focus on maximizing design freedom for vehicle styling are more likely to capture market share through innovation capacity, particularly during early stages of the product life cycle (Talke *et al.*, 2009). Given too little design reach relative to the market's desire for change and the brand's history of innovation, the product appears weak and stale: given too much reach, the customer reaction may be anxiety and discomfort(Berlyne, 1971). If the reach is in the wrong direction, because it either violates the brand's identity or strays from the benefits desired by the target market, the product may fail within the market (Hartley, 1996a).

On the automotive industry side, brand loyalty is a significant factor in customer purchase decisions. Brands such as BMW and Cadillac have taken more than 100 years to build a brand reputation; and oftentimes, in stated customer responses, brand is near or at the top in influencing purchase decisions (201, 2014c). By maintaining brand recognition, the equity of the brand may be leveraged for new products, thus influencing customer preference (Barney, 1991; Person & Snelders, 2010; Schmitt, 2012; Srinivasan *et al.*, 2006).

As a result, both design freedom and brand recognition are competing considerations during the design process for both the designer and the enterprise as a whole. Correctly balancing this tradeoff is paramount to realizing market success (Moulson & Sproles, 2000)—akin to musicians aiming to produce their next great hit while still sounding true to their unique musical style.

Automotive Design Process

The automotive design process may be conceptualized as a long sequence of depictions, each one becoming more detailed and realistic. The design may begin as just a verbal description (e.g., "The next generation Chevrolet Malibu, coming off engineering platform B, aimed at owners of midsize cars who want a versatile and

modern design at a moderate price."). Or it may start with some rough physical dimensions (e.g., overall length, width, and height, within specified bounds).

Over a number of months, the depiction gains specificity in terms of physical dimensions, features, and options. What began as a description in words and numbers eventually transitions, first to 2D images and eventually to 3D models and prototypes. In these latter stages many decisions are made that will affect the aesthetic appeal and projected image of the design, and consequently the emotional reaction of customers. While these decisions are ultimately based on the intuition of highly trained designers, there is a long history of attempts to influence these decisions with a more data-driven approach.

The most common approach has been to conduct theme studies where designs are shown to customers who then rate them on several dimensions (e.g., appeal, innovativeness, distinctiveness, sportiness), and also take part in focus groups. This approach has often fallen short because evidence counter to designer intuition is met with skepticism by the designers. Another issue is that design activity typically occurs for 6-12 months before any customer feedback is collected. This creates an environment where designers' preferred designs gain momentum and backing by management, and are subsequently less likely to be changed given preliminary customer data.

Aim of This Chapter

In this study, we measure how brand recognition and design freedom interact and trade off with each other for four automotive luxury vehicle brands—Audi, BMW, Cadillac, and Lexus. Luxury brands are chosen primarily due to strong brand affiliation in their market segment (Aaker & Keller, 1990; Mannering *et al.*, 1991). To make such measurements, we decompose both brand recognition and design freedom to a common set of styling design attributes—an approach supported by psychology

and design research suggesting styling design attributes such as 'aggressiveness' may be more representative of visceral human perceptions of design than geometric design variables such as '120 cm vehicle grill' (Norman *et al.*, 2003; Norman, 2004; Reid *et al.*, 2010b). By manipulating the values of these styling design attributes rather than geometric design variables, we can better trace relative changes in both design freedom and brand recognition.

Manipulation of these design attributes, however, still requires a mapping to the geometric design variables that the designer controls: We cannot choose the 'aggressiveness' level of the vehicle, but we can decide the width of the wheelbase. Accordingly, we build on a general methodology common in the design community—determining the values of design attributes as functions of the underlying geometric design variables using customer responses (Louridas, 1999; McWilliam & Dumas, 1997; Mulder-Nijkamp & Eggink, 2013). A key difference in our approach, however, is that we do not explicitly model the functional form of the nonlinear mapping between styling attributes and geometric variables. Instead, we crowdsource this mapping as a black-box function that is hypothesized to model the judgments of customers. This approach may be too simplistic—see, e.g., (MacDonald *et al.*, 2009)—but we adopt it here as a starting point to address our research question of measuring the balance between design freedom and brand recognition.

Our experimental procedure involved three steps: (1) Determination of styling attribute values for existing vehicles using a Markov chain derived for partial rankings over customer responses to 2D design representations; (2) Generation of new conceptual designs using morphable 3D design representations; and (3) Determination of design freedom and brand recognition via deviations from previous designs of both styling design attributes and geometric design variables, using a proposed design freedom distance metric and a conditional multinomial logit model. Customer responses and new concept designs were gathered using an online interactive survey

consisting of sequential design evaluation and design generation stages using both two-dimensional (2D) images and three-dimensional (3D) morphable vehicle models rendered in real time. Using the data from this experimental procedure, we quantitatively capture the relationship between design freedom and brand recognition on a brand-by-brand basis.

This research approach thus puts its entire emphasis on determining an accurate relationship between design freedom and brand recognition, at the expense of being unable to ask the reasons "why" this relationship exists. This is due to the use of nonparametric and nonlinear predictive models to assess design freedom; in which such models do not have a known functional form much less a known inverse. In other words, we do not know which sets of geometric design variables affect which perceptual design attributes, yet we know the value of its corresponding design freedom and brand recognition.

Significance of this Study

The results of this study show that there is indeed a tradeoff between brand recognition and design freedom according to the proposed design freedom metric. This tradeoff is predicted to significantly affect BMW and Cadillac the most, suggesting that these brands face greater challenges to maintain brand recognition while evolving the styling of future vehicles. The tradeoff is predicted to affect Audi and Lexus less, however these tradeoffs are less conclusive as both these brands are found to have low absolute brand recognition across customers surveyed throughout the world.

The main contribution of this work is an extension of previous *descriptive* investigations (Kreuzbauer & Malter, 2005; McCormack *et al.*, 2004b; Ranscombe *et al.*, 2012) of brand recognition and design freedom to a *predictive* investigation involving modeling of brand recognition and design freedom. While it is often qualitatively recognized that brand recognition and design freedom must trade off with each other,

we make an early effort to a quantitative measurement of this tradeoff.

This work does not seek to optimize the tradeoff between design freedom and brand recognition, which would require modeling decisions by a multitude of stakeholders—particularly designers, marketers, and strategic design managers. Instead, we posit that the present work can augment stakeholder intuition during the strategic design decision-making process.

Additional contributions include: (1) The combined use of multiple design representations for predictive modeling including styling attributes and more conventional geometric variables as previously studied (Ersal *et al.*, 2011; McWilliam & Dumas, 1997; Orsborn *et al.*, 2009a; Sylcott *et al.*, 2013b); (2) A hybrid combination of parametric models and non-parametric representations; (3) The use of realistic, morphable 3D modeling techniques in an interactive web-based environment, an approach gaining popularity in areas such as design co-creation (Ramanujan *et al.*, n.d.); (4) Filtering crowdsourced data on "brand-conscious" customers to filter data relevant for this study; and (5) Using the crowd as a "black box" for modeling an implicit nonlinear function distributed over a number of people.

Using the crowd as a "black box" is perhaps the most important methodological contribution. In particular, measuring styling has always been problematic because it is perhaps one of the most challenging problems from a statistical and modeling standpoint. In particular, a realistic design's styling, for example, a full 3D model of a vehicle, must be represented by more than 10,000 to 100,000 dimensions (e.g., a door handle has length, width, curvature dimensions, thickness, color, sheen, etc.). Building a function with unknown functional form that maps styling from this high-dimensional space to a single number is challenging. Instead, using crowdsourcing to 'discover' this function from the responses of a large number of people does not require making *a priori* functional form assumptions, similar to recent work on constructing implicit functions from kernel feature spaces (Ren & Papalambros, 2012b).

4.2 Related Work

Balancing between design freedom and brand recognition has been studied extensively in the product innovation and styling strategy literatures as well as the design research literature. From the strategic management and customer product innovation communities, we establish qualitative justifications for upholding design freedom and brand recognition. From the design community, we consider previous efforts toward measuring tradeoffs between design styling and other considerations, as well as methodologies towards eliciting customer preferences via various design representations.

Design Freedom and Brand Recognition

Several studies have considered the importance of design freedom from the perspective of organizational innovation capability, with a consensus that there is an optimal amount of deviation from previous designs (Hekkert *et al.*, 2003; Person *et al.*, 2008). Customers expect novelty in new product offerings (Martindale, 1990), yet such novelty must be bounded (Berlyne, 1971). Companies that follow a "design-driven" approach toward balancing this tradeoff via strategic design decisions have been shown empirically to capture larger market shares (Person *et al.*, 2008).

The effect of brand recognition on customer preferences has been studied in depth for new product offerings (Aaker & Keller, 1990). General conclusions from these studies are that brands are comprised of highly complex associations between within-brand products and features (Milburn & Childs, 2001; Ranawat *et al.*, 2012), as well as related people, places, and out-of-brand products (Keller, 2003). Particularly because automobiles fall under the category of "durables," namely, products where lifecycle use is important to the customer, brand recognition plays a very important role (Zeithaml, 1988). These conclusions are aligned with observations in the automotive sector, where brand has been shown to be one of the foremost contributors to customer

Figure 4.1: Example images shown to customers in the 2D representation portion of the experiment. These images were used to assess styling attribute values, as well as brand recognition. The images remained static (were not morphed by customers) during the experiment and did not contain brand logos.

preference (Mannering *et al.*, 1991; 201, 2014c).

The current chapter builds on recent results showing that the front fascia or "face" of the vehicle—the view looking directly at the front of the vehicle—is most closely associated with vehicle brand (Ranscombe *et al.*, 2012). Moreover, anecdotal evidence from experienced sources within the industry support this notion (Manoogian II, 2013). Accordingly, all stimuli used in this study consider the face view of vehicle designs.

Brand-Conscious Customers

Brand-conscious customers, able to correctly identify brand from unbranded vehicles, are used for filtering the data collected in the study. These brand-conscious customers are filtered, because data from customers unable to identify brand add noise to the construction of predictive models for brand recognition. Moreover, appealing to brand-conscious customers has been found to be important for premium brands such as those considered in this study (Aaker, 2009).

To identify brand-conscious customers, we filter out customers not able to correctly identify brands above a given threshold for designs that already exist in the

market (see below for filtering criteria). Recent literature in crowdsourcing research has shown that data from "experts" within a crowd, in this case "brand-conscious customers" within a crowd, may be aggregated to obtain an accurate 'crowd consensus vote' using simple algorithms such as majority vote (Sheng *et al.*, 2008b; Sheshadri & Lease, 2013b)). However, if such filtering on the "experts" in the crowd is not done, simple algorithms to aggregate customer input may result in heavily biased crowd-level evaluations (Burnap *et al.*, 2015b). In our current case, this may skew estimates of design freedom when trading off brand recognition. Such filtering of customer data to guide the design process has been similarly explored by using customers to interactively guide the creative aspect of early-stage design (Ind & Watt, 2006; Crilly *et al.*, 2004a)).

Design Representation

Design representation refers to the method that a design artifact is encoded by either a computer or a customer during one of many steps in the design process (Chandrasegaran *et al.*, 2013a). We make a distinction between the two as it has been shown that computer representations and human representations may be entirely different, resulting in the need to construct models and conduct experiments in the appropriate space (Tversky & Gati, 1978; Tversky & Hutchinson, 1986). Moreover, we consider three different forms of design representation: 2D and 3D model geometry; parametric and non-parametric geometry; and as a function of styling attributes and geometric variables.

2D and 3D Representations

Recent studies have shown that brand recognition is dependent on the fidelity of the design representation (Ranscombe *et al.*, 2012; Rasoulifar *et al.*, 2015). Informally, there is a level of realism to the design that must be achieved for customers to correctly

Figure 4.2: Example images shown to customers in the 2D representation portion of the experiment. These images were used to assess styling attribute values, as well as brand recognition. The images remained static (were not morphed by customers) during the experiment and did not contain brand logos.

identify vehicle brand (Orbay *et al.*, 2015a). We build on this notion by representing vehicle designs using the highest fidelity representation possible whether a 2D image or a 3D high polygon mesh, as shown in Figures 4.2 and 4.2, respectively.

Studies have also shown differences between 2D and 3D design representations regardless of fidelity. In particular, customer preferences assessed through conjoint analysis have been found to be inconsistent when contrasting the type of design representation (Bao *et al.*, 2014a; Reid *et al.*, 2013a; Toh & Miller, 2014a). The area of assessing the level of fidelity or abstraction to a given threshold for a customer's perception is still an active area research, including both 2D and 3D representations.

Parametric and Non-Parametric

Design representations may be also categorized as parametric or non-parametric. Parametric design representations have numerous applications via conjoint analysis using 2D silhouettes (Orsborn *et al.*, 2009a; Petiot & Dagher, 2010; Reid *et al.*, 2010b; Sylcott *et al.*, 2013b), gestalt quantification using 2D representations (Lugo *et al.*, 2015), and 3D interpolated Bezier curves (Ren *et al.*, 2013b; Tovares *et al.*, 2014a); however, perhaps the most realistic 3D interpolated Bezier curves come from design research done within the automotive industry (Kókai *et al.*, 2007a).

In the shape grammar literature, non-parametric design representations are used as basic constituent shape elements to generate larger and more complex forms. These include automotive applications (Orsborn *et al.*, 2006a; Orsborn & Cagan, 2009), some with focus on vehicle face details (McCormack *et al.*, 2004b) and vehicle side profiles (Bluntzer *et al.*, 2014; Pugliese & Cagan, 2002a; Yannou *et al.*, 2008a). Such shape grammar techniques are applicable to generation of 3D design representations, for example, with fluid channel layouts (Hooshmand & Campbell, 2014).

The representation approach here is qualitatively similar to the shape grammar approach in that it employs a design generation process where an agent creates new designs, but it is limited in scope when contrasting the corresponding design spaces. In particular, shape grammars are able to generate a much larger set of possible designs as defined by the Cartesian product of grammar enumeration, whereas the design generation considered in this study is limited to the convex hull defined by the morphing bounds on the 3D design representations.

In this study, we cast the 3D design representation as a set of geometric features that morph not strictly related via a mathematical function, nor non-parametrically such as volumetric deformation (Tiwari *et al.*, 2014), but instead requiring pre-defined input, say from professional vehicle designers (Manoogian II, 2013). This results in a hybrid of the parametric and non-parametric design representations, where a number of geometric features morphs the 3D design via Laplacian deformation of its constituent polygonal mesh (Botsch & Sorkine, 2008). Note that we use morphable 3D models but only static images for the 2D design representations.

Visceral Attributes and Geometric Variables

While geometric variables via 2D and 3D representations, parametric or non-parametric, capture the physical form of the design as a computer may interpret it, human perceptions are better suited to a different representation (Coates, 2003;

Norman, 2004). In particular, design attributes such as 'Friendly' versus 'Aggressive' have been posited to represent human perceptual understanding better than variables such as '130 cm long airdam' (Norman, 2004).

To develop analytical decision-making models(Papalambros & Wilde, 2000), we further assume that the attributes themselves are functions of geometric design variables. Styling attributes are likely nonlinear functions of geometric variables, e.g., slight geometric changes in the edges between a smile and a frown may make large differences in an attribute such as 'happiness' (201, 2014c). By gathering customer responses within the space of design attributes versus design variables, we are operating at a level analogous to similarity models in the psychological literature (Tversky & Hutchinson, 1986).

Quantitative Models of Product Styling

Previous research in quantitative modeling of styling and aesthetics has often come from the marketing community, where conjoint analysis has proven valuable (Green *et al.*, 1981). This modeling technique takes a number of variables representing the design's form as input and elicits customer preferences across a set of discrete points within the design space.

The design community has similarly used conjoint analysis to model styling form in efforts to optimize customer preferences in decision-based design (Chen *et al.*, 2013; Hazelrigg, 1998; Papalambros, 2002). Relevant examples of such applications include 2D vehicle side view silhouettes (Orsborn *et al.*, 2009a; Reid *et al.*, 2010b) and 2D vehicle faces (Petiot & Dagher, 2010). Recently, 3D vehicles studies such as perceived safety (Ren *et al.*, 2013b) and vehicle interiors (Poirson *et al.*, 2013a), as well as virtual reality studies (Tovares *et al.*, 2014a) have been investigated. Some applications have used nonlinear conjoint models such as explicit feature mappings (Fuge *et al.*, 2013) and implicit feature mappings (Ren *et al.*, 2013b). Additional 3D extensions include

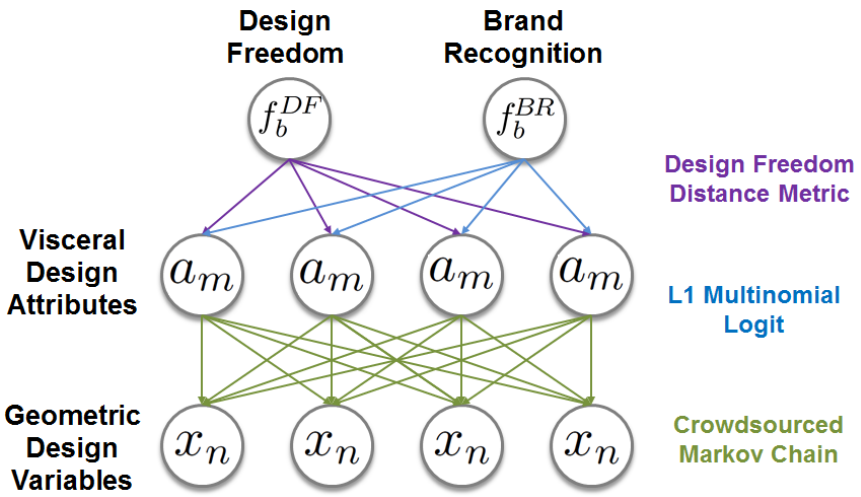


Figure 4.3: Dependencies between design freedom and brand recognition, design attributes, and design variables. Note that while design freedom and brand recognition are explicit linear functions of design attributes, design attributes are nonlinear functions of geometric design variables implicit in the customer perceptions of vehicles. On the right hand side, we denote the functional form of the associated dependencies.

the use of hierarchical geometric representations that may be used for salient feature extraction (Orbay *et al.*, 2015b).

4.3 Problem Formulation

We formally define brand recognition and design freedom, and the manner in which the two are measured. We additionally define how customer responses to conceptual designs are aggregated to assess the overall crowd consensus to changes in conceptual designs.

Let $f_b^{DF} : \mathscr{A} \to \mathbb{R}$ and $f_b^{BR} : \mathscr{A} \to [0, 1]$ denote design freedom and brand recognition, respectively, in which $\mathscr{A} = \{\mathbf{a} = [a_1, \ldots, a_M] : a_m \in \mathbb{R}\}$ is the space of styling attribute vectors, and $\mathbb{R}$ is the real space. Note that as discussed in the background section, this definition assumes the styling design attributes are a common set of inputs to both design freedom $f_b^{DF}(\mathbf{a})$ and brand recognition $f_b^{BR}(\mathbf{a})$, and that both are defined over the set of existing and conceptual designs $\mathbf{x} \in \{\mathbf{x} = [x_1, \ldots, x_N] : a_n \in \mathbb{R}\}$

for an associated brand $b = 1 \ldots B$.

These design attributes ${a_m}_1^M$ are defined as the building blocks of customer perceptual representation of design styling, following the idea of how human perception is chunked (Norman, 2004). Informally, humans conceptualize a vehicle using terms such as 'sportiness' rather than a large number of geometric design variables that constitute sportiness such as 'length of upper air dam.'

The design attributes must be empirically manipulated to measure relative changes across brand recognition and design freedom. Accordingly, we parameterize the design attributes as a nonlinear function of a set of predefined geometric design variables denoted {x_n }_1^N. This parameterization attempts to capture the notion that changing a given design variable may affect multiple attributes at the same time in a complex manner.

The dependencies of design freedom and brand recognition, design attributes, and design variables are shown in Figure 4.2. We next define the functional form of each dependency. In particular, we detail the mathematical relationship between (1) design freedom and design attributes, (2) brand recognition and design attributes, and (3) design attributes and design variables.

4.3.1 L1 Multinomial Logit for Brand Recognition

We define brand recognition as a linear combination of design attributes, in which the attributes maximally discriminate between brands. To determine the linear coefficients to predict brand, we assume a multinomial logistic regression functional form, conditioned only on brand-conscious customers and regularized using the L1-norm, as given in Equation (4.1).

$$f_b^{BR}(\mathbf{a}) = \frac{e^{\mathbf{w}_b^T \mathbf{a}}}{\sum_{b=1}^{B} e^{\mathbf{w}_b^T \mathbf{a}}} + |\mathbf{w}_b|_1 \tag{4.1}$$

To train the coefficients $\mathbf{w}_b$ of this model, we use a quasi-Newton optimization

algorithm (l-BFGS) to maximize the penalized multinomial likelihood (Papalambros & Wilde, 2000). Note that we use here the notation for coefficients from the machine learning literature; these coefficients are also often denoted with the symbols β in marketing and θ in statistics. The data are conditioned using a hard threshold, where a brand-conscious customer must achieve greater than T percentage correct recognition of brands across a set of existing designs.

4.3.2 Design Freedom Distance Metric

Design freedom is the leeway designers have to generate conceptual designs while accounting for many implicit and explicit constraints (Hartley, 1996a). To capture this leeway, we adopt the information processing flow in Crilly et al. (Crilly *et al.*, 2004a) by assuming that the communication from designer to customer is conveyed through information of multiple modes—in our case a vector of design attributes and vector of geometric values representing the design artifact.

With this design representation of multiple modes, we define design freedom as a distance from existing designs to a new conceptual design both across design attributes and across geometric variables. This design freedom distance is mathematically captured using a distance metrics; yet while various metrics have been previously used for engineering specifications (Simpson *et al.*, 1998), representations such as abstract knowledge databases (Chandrasegaran *et al.*, 2013a), and text (Fu *et al.*, 2013b), these metrics do not accommodate various stakeholder inputs as specifically needed in the current chapter.

We thus propose a distance metric between two designs α and β for brand b as given in Eq. (4.2). This metric is used to assign a scalar value that captures both geometric and perceptual styling differences between designs.

$$||f_b^{DF,\alpha} - f_b^{DF,\beta}|| = \sum_{m=1}^{M} \mathbb{I}_{w_{b,m\neq 0}} \left[\lambda_1 \left(a_m^{(\alpha)} - a_m^{(\beta)} \right)^2 + \lambda_2 \sum_{n=1}^{N} r_{b,nm} \left(x_n^{(\alpha)} - x_n^{(\beta)} \right)^2 \right] \quad (4.2)$$

where,

$$\begin{aligned}
a_m &= \text{design attributes measured using 2D representation} \\
x_n &= \text{geometric design variables common to both 2D and 3D representation} \\
\lambda_1 &= \text{importance/normalizing operator of design attributes} \\
\lambda_2 &= \text{importance/normalizing operator of geometric design variables} \\
\mathbb{I}_{w_{b,m\neq 0}} &= \text{indicator function if attribute } m \text{ is important for brand } b \\
r_{b,nm} &= \text{sensitivity of attribute } m \text{ to variable } n \text{ for brand } b
\end{aligned}$$

This distance metric captures stakeholder considerations to the overall design freedom in two ways: First, design freedom *implicit* in the mind of the customer is captured using $r_{(b,nm)}$ and $\mathbb{I}_{w_{b,m\neq 0}}$, both of which are assessed using the customer crowd. Informally, these values capture the notion that differences between two designs exist with both geometric and perceptual representations in the mind of the customer.

Second, design freedom explicit from stakeholders within the producing organization are captured using λ_1 and λ_2, which may represent, say, relative influences of the marketing and engineering departments, respectively. Informally, we use these operators to tune how important it is to maintain an attribute like "aggressiveness" for a marketing campaign, or a certain geometric shape for vehicle aerodynamics. Accordingly, these operators are specific to the brand being considered.

Using this distance metric, overall design freedom is assessed as the distance from the current design in Model Year 2014 (MY2014) to a proposed design $(\mathbf{x}', \mathbf{a}')$. De-

noting the current design $(\mathbf{x}^0, \mathbf{a}^0)$, design freedom for the proposed design is given by Eq. (4.3) using vector notation for brevity.

$$\begin{aligned} f_b^{DF}(\mathbf{x}', \mathbf{a}') &= ||f_b^{DF}(\mathbf{x}', \mathbf{a}') - f_b^{DF}(\mathbf{x}^0, \mathbf{a}^0)|| \\ &= \lambda_1(\mathbf{a}^{'} - \mathbf{a}^0)^T \text{diag}\left[\mathbb{I}_{\mathbf{w}_{b\neq 0}}\right](\mathbf{a}^{'} - \mathbf{a}^0) \\ &+ \lambda_2(\mathbf{x}^{'} - \mathbf{x}^0)^T \text{diag}\left[\mathbf{R}\mathbb{I}_{\mathbf{w}_{b\neq 0}}\right](\mathbf{x}^{'} - \mathbf{x}^0) \end{aligned} \tag{4.3}$$

where,

$$\begin{aligned} \mathbb{I}_{\mathbf{w}_{b\neq 0}} &= Mx1 \text{ vector of indicator functions for brand } b \\ \mathbf{R} &= NxM \text{ matrix of attribute} - \text{variable sensitivities} \\ \text{diag}\left[\cdot\right] &= \text{operator to transform vectors to diagonal matrices} \end{aligned}$$

To calculate the sensitivities of design attributes to design variables $r_{(b,nm)}$, we conduct a one-sided t-test between the baseline design variable x_n^0 and the morphed $x_n^{'}$ from customer responses for a given attribute m and brand b. This hypothesis test sets the $r_{(b,nm)} = 1$ if the p-value for the t-test is less than 0.05, and $r_{(b,nm)} = 0$ otherwise. The values of the indicator function $\mathbb{I}_{w_{b,m\neq 0}}$ are calculated by assigning the value 1 to all non-zero elements of the corresponding weight vector described in Section 3.1. This weight vector is already sparse due to L1 regularization, and is thus suited to picking out attributes that most contribute to the brand styling (Ranawat *et al.*, 2012).

4.3.3 Crowdsourced Markov Chain for Design Attributes

Our next goal is to develop a method of obtaining the attribute values for each design, for example, "this vehicle is 0.7 out of 1.0 for aggressiveness, 0.2 out of 1.0 for distinctiveness," and so on. This method is a function that maps a design's geometric

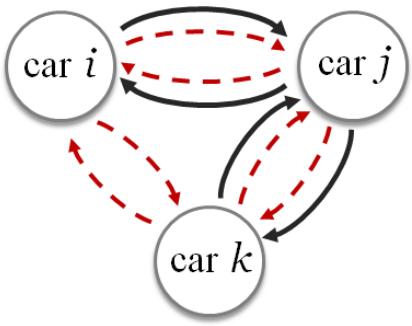


Figure 4.4: Diagram of Markov chain used to aggregate customer responses in the form of partial rankings of cars to obtain design attribute values for each brand. Black arrows show non-zero transition probabilities from the raw transition matrix, while red dashed arrows show perturbation probabilities added to ensure a unique stationary distribution.

variables x to the design's corresponding attributes a.

Within the design community, this function has conventionally been approximated by explicitly assuming a functional form, such as the linear logit model often used in design utility theory treatments, followed by estimating part-worth coefficients of the assumed model. However, this function is likely highly nonlinear, particularly when dealing with high-dimensional representations required for realistic design stimuli.

Here we take a different approach by assuming that the nonlinear function relating design attributes to design variables is implicitly captured by the responses of brand-conscious customers. By crowdsourcing the attribute values of the designs—asking a crowd of customers to evaluate designs over attributes—we avoid needing to make a priori assumptions regarding this complex nonlinear functional form explicitly. This has advantages as we are capturing a function that may exist in a more expressive function space, allowing complex modeling of nonlinear interactions. Moreover, we avoid the need of explicit mathematical representation of geometric variables, given that realistic 3D vehicle polygon meshes may contain more than 100,000 vertices.

Under this approach, there are several ways to extract attribute values of designs from the evaluations provided by the crowd. We choose to extract these values using only relative comparisons between the set of designs, avoiding the notion of a non-

relative scale, i.e., "what would it mean to give a design a 0.4 out of 1.0 'aggressive' score without seeing the entire set of designs, and how could we ensure everyone used the same scale?"

In particular, we ask the crowd to evaluate the attributes of designs as a ranking between just a few designs at a time. Formally, the responses r_{c1}^{C} made by customers $c = 1 \ldots C$, in which each evaluation is in the form of a partial ranking for a single design attribute. Partial rankings without ties are chosen as more intuitive for human evaluation (Gonzalez & Nelson, 1996).

To obtain attribute values using this set of evaluation responses from the crowd, i.e., to aggregate these partial rankings into numbers for each attribute and for each design, we derive a Markov chain solved using a modified version of PageRank (Brin & Page, 1998) as given in Equations (4.4), (4.5), and (4.6). Informally, this Markov chain treats the ranking of all designs for a specific attribute as a set of "states," for car designs to jump between. Every time a car is ranked above another, that car pair jumps to the higher ranked state. The set of states that correspond to the maximal number of correctly ranked cars is called the "stationary distribution." Finding this desired final ranking of states requires an iterative optimization procedure.

This iterative procedure is characterized by the Markov chain jumping around to different states as shown in Figure 4.3.2. This jumping action is governed by a transition probability from one state to another, and in our case those transition probabilities depend on partial rankings. The converged stationary probability distribution of the Markov chain is then used as the value of the attribute. Specifically, we define the attribute value as the probability that the car is ranked higher than other cars, thus the attribute value of a car equals its average percentage of the time that it is ranked higher than other competing cars. Following the jumping analogy, if a car is more likely to be ranked higher than others, then the agent will jump into that state more often.

More formally, the transition probability $P_{ij}, i, j = 1, \ldots, N$ from the state representing car i to the state representing car j is defined as the frequency that car j is ranked higher than car i in all partial ranks that contain car i. If the transition probability P_{ij} is large, we define car j as being more likely to have greater relative attribute value than car i. We denote the transition probability matrix as $\mathbf{P} = (P_{ij})$, hereafter referred to as the raw transition probability matrix. The stationary distribution π of a Markov chain is a distribution vector unchanged after the operation of transition matrix $\mathbf{P}$, as given in Eq. (4.4).

$$\begin{aligned} \pi &= \pi\mathbf{P} \\ \pi &= (\pi_1, \pi_2, \ldots, \pi_N) \\ \pi_i &\geq 0 \text{ and } \sum_{i=1}^{N} \pi_i = 1 \end{aligned} \tag{4.4}$$

Consistent with Markov chain theory, there is no guarantee that the raw transition probability matrix $\mathbf{P}$ will have unique stationary distribution (Ross, 1996) without some strong assumptions. To achieve uniqueness in the resulting distribution, we make two extensions to convert the raw transition matrix $\mathbf{P}$ to a stochastic, irreducible, and aperiodic matrix (Brin & Page, 1998).

Extension 1.

The rows in $\mathbf{P}$ containing only 0's are replaced with $\frac{1}{N}\mathbf{e}^T$, where $\mathbf{e}^T$ is a column vector consisting of 1's, and T denotes the transpose operator. This adjustment results in a stochastic matrix denoted $\mathbf{S}$ as given in Eq. (4.5).

$$\mathbf{S} = \mathbf{P} + \mathbf{Q}\left(\frac{1}{N}\mathbf{e}^T\right) \tag{4.5}$$

$$Q_i = \begin{cases} 1 & \text{if } P_i = \mathbf{0} \\ 0 & \text{otherwise} \end{cases}$$

Extension 2.

To convert $\mathbf{S}$ into an irreducible and aperiodic matrix $\mathbf{G}$, we use Eq. (4.6).

$$\mathbf{G} = \gamma\mathbf{S} + (1-\gamma)\frac{1}{N}\mathbf{e}\mathbf{e}^T \tag{4.6}$$

where γ is a scalar between 0 and 1 controlling the intensity of the perturbation that ensures uniqueness.

With these extensions, a unique stationary distribution exists for $\mathbf{G}$. From Eq. (4.6), the stationary distribution vector π can be obtained by calculating the eigenvectors of $\mathbf{G}$ or by iteratively calculating $\pi^{(k+1)} = \pi^{(k)}\mathbf{G}, k = 1, 2, \ldots$ until convergence. To calculate the values of attributes $\mathbf{a}_b$ for brand b based on the set of all partial rankings from customer responses r_{c1}^C, we define the attribute value for car i as $\pi_m, m = 1, 2, \ldots, M$.

4.4 Experiment

We conducted two experiments to measure how brand recognition changed as design freedom increased. Experiment 1 assessed brand recognition using 2D images of current MY2014 vehicle designs, followed by generation of new morphed concept designs using 3D morphable models. Experiment 2 assessed brand recognition using 2D images of both the MY2014 vehicle designs and the new morphed concept designs.

The data collected from the MY2014 baseline designs allowed us to measure current brand recognition for each brand, as well as develop a predictive model for brand recognition as a function of design attributes. The data collected from the morphed concept designs allowed us to measure brand recognition at various values of design freedom.

Customers

We gathered a total of 315 customers through the crowdsourcing platform Amazon Mechanical Turk (201, 2014a). As online crowdsourcing has been empirically shown to be a noisy process, partially due to various motivations of customers (Gerth *et al.*, 2012; Panchal, 2015b; Pilz & Gewald, 2013; Sheshadri & Lease, 2013b), we filtered out data from customers using two data processing steps to ensure data fidelity.

First, customers that simply "clicked through" the survey were filtered out by requiring their average time on the 2D portion of the site to be greater than 6 seconds per ranking. Second, a brand recognition accuracy threshold of 30% was chosen to filter out customers who were not "brand-conscious" as justified in Section 2. For reference, the average brand recognition accuracy for the unfiltered crowd was 32.78%. Brand recognition accuracy was treated as a constant variable across the entire survey, and all data were filtered out for a given participant if he or she did not fall above the threshold.

After filtering mechanisms, 139 customers were retained from a total of 315 customers gathered over two experiments, as described in the experimental procedure. In particular, experiments gathered 196 and 119 customers, of which 96 and 43 remained after filtering, respectively.

Brand	Compact	Midsize	Fullsize	Crossover	SUV
Audi	A4	A6	A8	Q5	Q7
BMW	3-series	5-series	7-series	X3	X5
Cadillac	ATS	CTS	XTS	SRX	Escalade
Lexus	IS	GS	LS	RX	GX

Table 4.1: Description of the four vehicle manufacturer brands and five associated vehicle classes used in this study.

Vehicle Brands and Models

The brands chosen were Audi, BMW, Cadillac, and Lexus ($B = 4$), due to their relative similarities over a targeted market segment of luxury vehicles, as well as similarity of product offerings across vehicle classes. For each brand, five models were chosen from MY2014 corresponding to five vehicle classes as given in Table 4.4.

2D Images and 3D Morphable Vehicle Models

Images of the vehicle face were sourced from an online vendor (201, 2014b). The face image has been shown to be more correlated with brand recognition than side view or rear vehicle view (Ranscombe *et al.*, 2012). Each image consisted of a white vehicle on a white background to minimize confounding interactions from color as shown in Figure 4.2. Moreover, the brand logo was removed for each vehicle image in order to focus customer responses just on styling as in (Fu & Kara, n.d.).

Four morphable 3D models, one for each brand, were created as shown in Figure 2. Morphing was pre-computed offline using Laplacian deformation and volumetric-based mesh deformation techniques (Botsch & Sorkine, 2008). The models were then imported into the web-based survey using the browser-based WebGL renderer, allowing real-time and realistic deformation via client-side graphics processing unit interpolation.

Low Attribute	High Attribute
Awkward	Well Proportioned
Weak	Powerful
Conservative	Sporty
Basic	Luxurious
Conventional	Distinctive

Low Attribute	High Attribute
Passive	Active
Traditional	Innovative
Understated	Expressive
Friendly	Aggressive
Mature	Youthful

Table 4.2: Description of the four vehicle manufacturer brands and five associated vehicle classes used in this study.

Design Attributes

As discussed above, design attributes link brand recognition with design freedom. We selected ten design attributes given in Table 4.4 based on input from actual design teams in the automotive industry (201, 2014c).

Experimental Procedure

We conducted two experiments: Experiment 1 gathered attribute values and brand recognition accuracies via partial rankings of 2D images of MY2014 baseline designs. This was followed by generation of new morphed concept designs using a 3D morphable model. Experiment 2 similarly gathered attribute values and brand recognition accuracies, except this time using 2D images of both the MY2014 baseline designs mixed with 2D images of the 3D morphed concept designs from Experiment 1. The overall procedure is given in Figure 4.4, and was as follows:

Experiment 1

Participants were first directed to an introduction page, where they were given instructions on ranking vehicles according to a semantic differential. This semantic differential consisted of only one of the ten attributes from low to high value or vice versa to act as a counterbalance for ordering biases. Over the entire interactive survey, a participant was always given the same semantic differential to reduce participant burden.

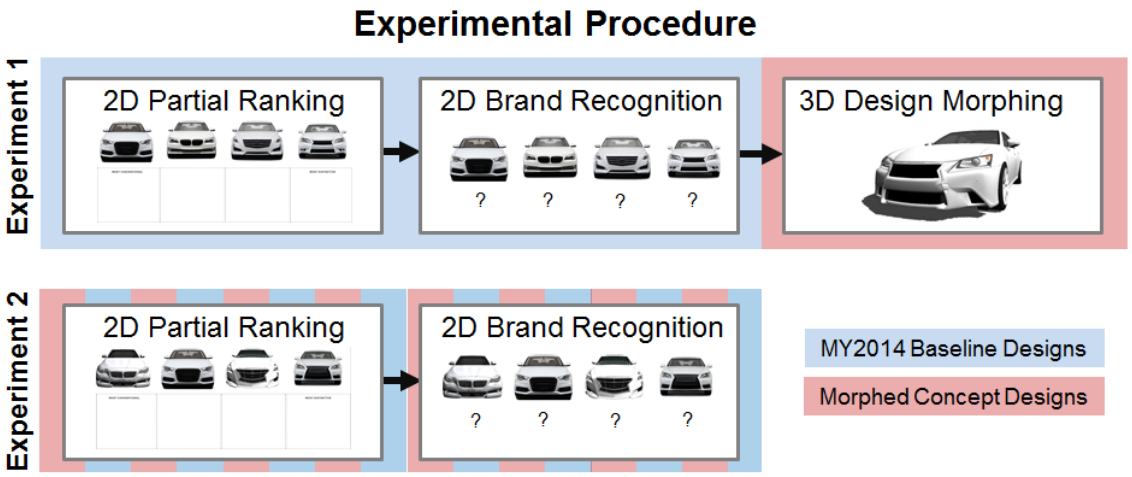


Figure 4.5: Overview of the experimental procedure for both Experiment 1 and Experiment 2. Experiment 1 asked participants to give partial rankings of current MY2014 baseline designs for a given design attribute, followed by asking which brand each of the images corresponded to. Participants were then asked to morph a 3D design to create new concept designs given the same design attribute. Experiment 2 asked a different set of participants to give partial rankings of current MY2014 baseline designs mixed with images of the morphed concept designs from Experiment 1. Similarly, participants were then asked which each brand the images corresponded to.

Next, participants were directed to the 2D design ranking page, with the four vehicles in a top row and four outlined placeholders in a bottom row. Instructions on the page were given to drag-and-drop the four MY2014 baseline designs from the top row to the bottom row using the mouse, including possibility of reordering the partial ranking. Upon clicking the "Submit" button for the partial ranking, participants were then asked to choose the brand of each MY2014 baseline design using a drop-down menu with 34 possible options (e.g., Audi, Volvo, Toyota).

After participants chose a recognized brand for each of the vehicles, they were allowed to continue to the next partial ranking. After participants completed five partial rankings on the 2D portion of the site, they were directed to the 3D portion of the site for generating new designs. In this portion, each participant was given a randomly chosen 3D model in the midsize vehicle segment from the four brands.

Participants were then asked to maximize the same design attribute as their semantic differential from the 2D portion of the site by morphing the 3D design using four sliders. They were able to rotate the 3D vehicle model to assess the gestalt of the face. Upon submitting their chosen 3D design, participants were then directed to a short survey in which they were asked basic demographic information as well as task relevant information.

Experiment 2

A new set of participants was asked to give partial rankings for a randomly assigned design attribute, but now with both 2D images of MY2014 baseline designs mixed with 2D images of face views of the 3D morphed concept designs from Experiment 1. Recall that this mixture of MY2014 baseline designs and morphed concept designs is necessary to get relative attribute values using the partial ranking Markov chain method derived earlier. A total of 52 possible 2D images was shown to participants, 20 from the original MY2014 baseline designs given in Table 4.4, and 32 from 3D designs morphed concept designs from Experiment 1.

Data Analysis

We give a diagram in Figure 4.4 of the data analysis using the methods detailed and developed in Section 4.3, and list this methods flow here: We aggregated the partial rankings from each brand-conscious participant using the method described earlier to obtain the design attribute values for each new conceptual design. These design attributes were used to build a model of brand recognition. The filtered data included participants from 2D images of both morphed and non-morphed designs due to the relative values obtained using the partial ranking aggregation method.

Brand recognition was assessed by calculating the number of correct responses to th set of 32 morphed conceptual designs over the total number of times that that

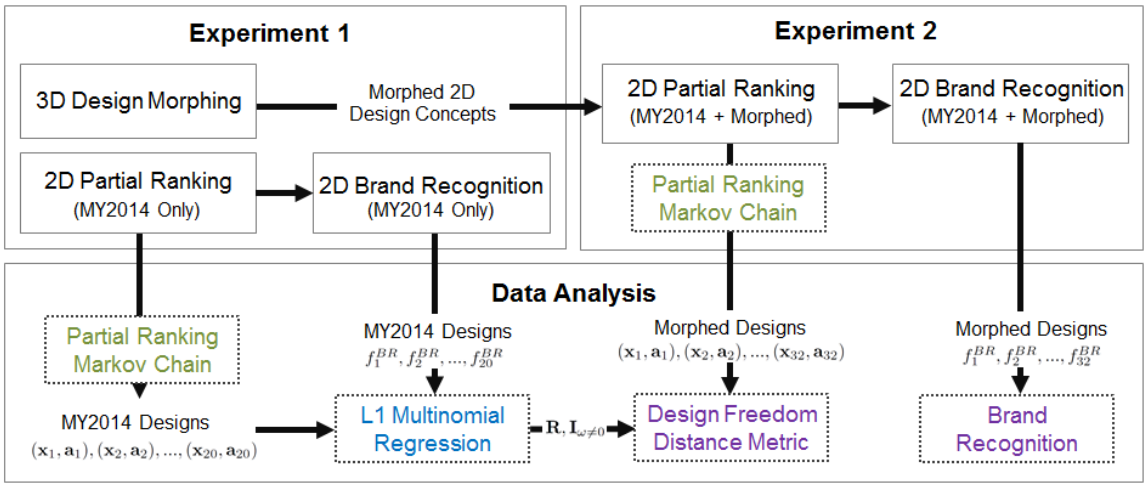


Figure 4.6: Diagram of the data flow and methods used in the data analysis of the experiment. As described earlier and shown in Figure 5, Experiment 1 provides the Partial Ranking Markov Chain and L1 Multinomial Regression with data from only MY2014 vehicle designs, thus provided the attribute-variable sensitivities $\mathbf{R}$ and brand-attribute sensitivities $\mathbb{I}_{(\omega \neq 0)}$. Experiment 2 provides the Partial Ranking Markov Chain with combined MY2014 and morphed vehicle designs, of which only morphed design attributes and variables are passed on to the Design Freedom Distance Metric. The values of design freedom for each morphed design are then compared with their corresponding brand recognition to obtain the desired slope on a brand-by-brand basis.

particular conceptual design showed up in the partial rankings. Design freedom was calculated using the metric described above. The operators λ_1 and λ_2 are chosen to scale the design freedom by subtracting the mean and dividing the standard deviation of each brand's design variables and design attributes, respectively, resulting in a normalized design freedom. This operation was chosen on a brand-by-brand basis as this did not change the brand recognition versus design freedom distributions.

4.5 Results

Four plots depicting the empirical relationships between brand recognition and design freedom for each manufacturer are given in Figure 4.4. Each plot includes a trend line obtained using Thiel-Sen robust linear regression to assess, to first order

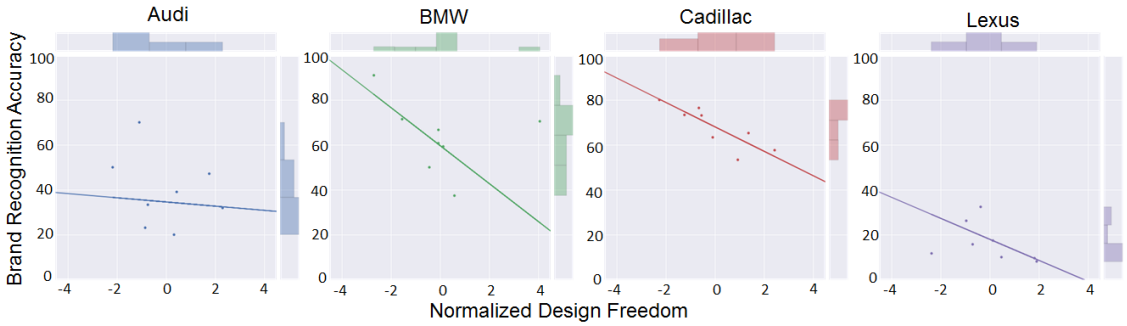


Figure 4.7: Brand recognition versus design freedom for the four vehicle brands in this study over 2D images taken of the conceptual designs generated during the 3D portion of the experiment. Brand recognition accuracy is defined as the percentage of time a brand-conscious customer—a customer who correctly identified more than 30% of the MY2014 baseline vehicle brands—was able to correctly recognize a new morphed design.

Brand	**Slope of Brand Recognition vs Design Freedom**	**Median Absolute Deviation**
Audi	-0.009	0.302
BMW	-0.085	0.074
Cadillac	-0.054	0.134
Lexus	-0.047	0.426

Table 4.3: Slope coefficients of Thiel-Sen robust linear model fit to brand recognition vs. design freedom for the four brands in this study.

only, how fast brand recognition decreases as design freedom is increased (Sen, 1968). The slopes of each of these trend lines is given in Table 4.5. As given by the median absolute deviation, linear relations for Audi and Lexus are not very meaningful but one can still discern a trend. Histograms showing the marginal distributions are also plotted in Figure 4.4 to convey the relative coverage of the data for each brand.

The brand recognition versus design freedom slope for each of the four manufacturers is negative, a result obtained entirely from the data, confirming intuition that increasing design freedom results in decreased brand recognition. From these slopes, we can see that BMW and Cadillac have the quickest loss of brand recognition with increasing design freedom. These results suggest that designers at BMW have much less leeway in their freedom to create future design concepts without sacrificing brand

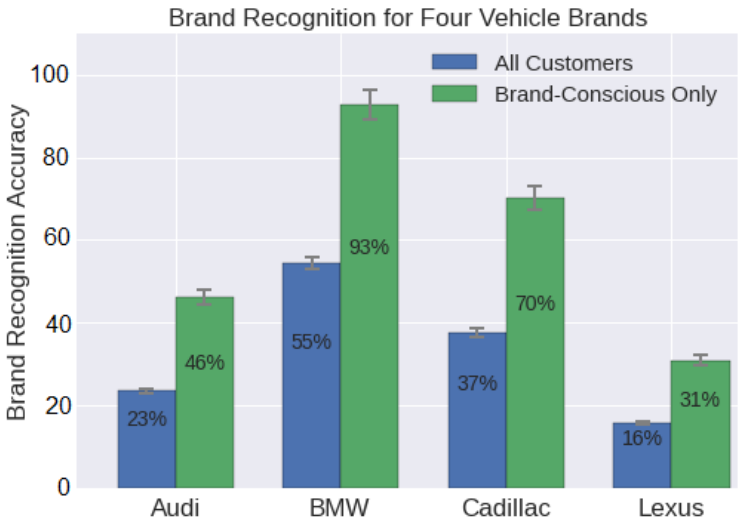


Figure 4.8: Brand recognition for the four vehicle brands in this study. Brand-conscious customers refer to those customers who could correctly identify at least on average 30% the brands of baseline (MY2014) designs.

image and heritage. Cadillac is second in this ordering, yet has significantly less sharp of a slope, suggesting that designers at Cadillac are not as constrained as designers at BMW.

Lexus and Audi are shown to be third and fourth in this ranking; however, both of these manufacturers have results that are less meaningful due to both poor linear fit as given by Table 4.5, as well as low overall brand recognition. In particular, Figure 8 shows the overall brand recognition accuracy across the four brands for both brand-conscious customers and non-brand conscious customers. We observe that BMW and Cadillac have the most recognizable brand, justified as the 'All Customers' data consist of over 5428 brand identifications from a pool of 315 customers. Audi and Lexus were found to have the lowest brand recognition, both among brand-conscious customers and non-filtered customers.

Application to Industry

The study was inspired by working with real automotive design teams, and direct practical implementation seems likely. One such implementation may be a tool to

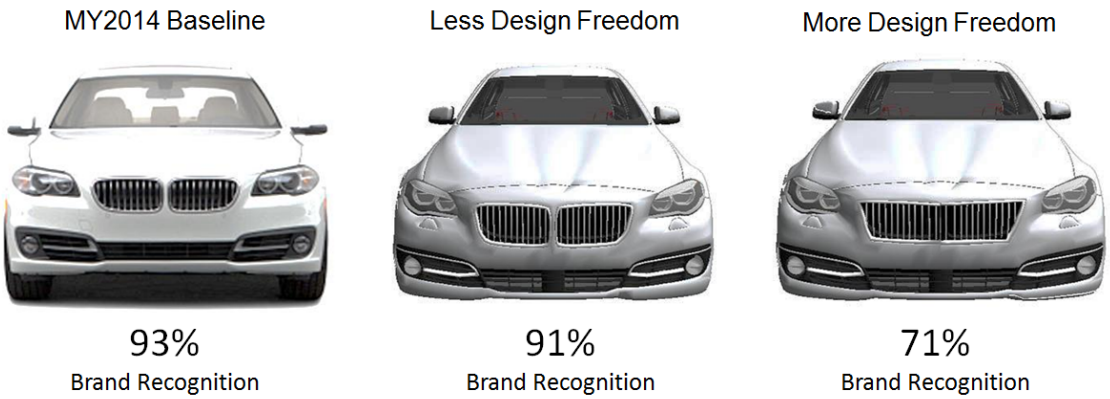


Figure 4.9: Example application to industry of the approach and results of this study. Three representations are given corresponding to the MY2014 Baseline BMW 5 Series, the morphed BMW 5 Series with the least design freedom from the baseline, and the morphed BMW 5 Series with the most design freedom from the baseline according to the data. Note that the MY2014 baseline is a 2D image, while the two morphed vehicles are images of the 3D morphing model.

generate "thought seeds" to act as inspiration for new design concepts. Such thought seeds may be used at very early stages of the design process in an effort to inspire creativity in directions that are most likely successful in the marketplace. Another implementation may be a check for promising design concepts such that they may be steered away from areas of "too much brand recognition" and not enough innovation and appeal, or on the other hand, "too much design freedom" and not enough resemblance to the brand and the current product family.

A future implementation could be a decision support tool for product researchers and strategic design managers to document explicitly which visceral design attributes and geometric design variables have the most leeway when creating a future design. As an example, we show in Figure 4.5 a baseline BMW 5 Series, along with two morphed BMW 5 Series with the least and most design reach from the baseline according to the data. For this example, we can see that the "kidney bean" grill significantly affects the relationship between design freedom and brand recognition; however, this was intuited a priori and is only able to be confirmed by the present works methodology—i.e., we

did not find this result introspectively—see the Limitations section for more discussion of this lack of ability to ask "why" design freedom trades off with brand recognition.

Such tools could augment the experience and intuition of designers and strategic design managers using real-time feedback from a targeted crowd of customers. This tool could be combined with more advanced 3D design and semantic representations (Yumer *et al.*, 2015c), gamification of real-time crowd feedback (Ren *et al.*, 2015a), and advances in virtual and augmented reality technologies (Ramani *et al.*, 2014; Shankar & Rai, 2014; Tovares *et al.*, 2014a; Faas *et al.*, 2014). Combinations of these recent technologies, all characterized by having a human-in-the-loop, would likely improve the outcomes of such efforts.

4.6 Summary

Design freedom and brand recognition are considerations that were measured for four vehicle manufacturers—Audi, BMW, Cadillac, and Lexus—since balancing between these two considerations has been shown to influence consumer purchase decisions significantly. An experiment was conducted measuring change in ten styling attributes common to both design freedom and brand recognition for automotive designs, soliciting customer responses to vehicle designs created interactively using 2D and 3D design representations. Results show that, while brand recognition is highly dependent on the particular vehicle manufacturer, measuring tradeoffs between design freedom and brand recognition using predictive models can augment human intuition in making strategic design decisions.

CHAPTER V

A Representation to Assess Evaluations and Preferences

5.1 Context: A "perfect" product form design tool?

When developing product form for a new design concept, human designers use a mental representation of possible concept designs that implicitly defines the "true" conceptual design space (Rosenman & Gero, 1993; Goldschmidt, 1997; Gero & Maher, 2013; Crilly *et al.*, 2004b) and is restricted only by the designer's cognitive skills. This true design space is searched using human creativity and experience (Hartley, 1996b; Cross, 2004b; Eckert & Stacey, 2000), with a search process that is both *flexible*, i.e., moving from one design to another happens naturally and fluidly, and *realistic*, i.e., product form representations mirror their eventual embodiment or, if the representation is an abstraction such as a sketch, convey sufficient information to capture the eventual design embodiment (Kavakli & Gero, 2001b).

Quantitative design methods use explicit mathematical representations of the design space, constituted of formalized elements such as vectors (Orsborn *et al.*, 2009b; Reid *et al.*, 2010a; Petiot *et al.*, 2009), trees (Orbay *et al.*, 2015c), graphs (Bayrak *et al.*, 2013b; Zhang & Rai, 2014), control points or handles for 2D pixels (Toh & Miller, 2014b; Bao *et al.*, 2014b) or 3D voxels (Yumer *et al.*, 2015b; Mukherjee *et al.*,

2014; Kang & Tucker, 2015; Ren *et al.*, 2013a; Tovares *et al.*, 2014b), and 2D (Pugliese & Cagan, 2002b; McCormack *et al.*, 2004a; Orsborn *et al.*, 2006b; Yannou *et al.*, 2008b) and 3D shape grammars (Perez Mata *et al.*, 2015). Each unique combination of elements and their values represents a design. These explicit formal representations tend to be either flexible but of limited realism, due to being low dimensional (e.g., a silhouette), or realistic but of limited flexibility, due to being high dimensional but only flexible in the local design space (e.g., 3D polygon mesh morphed by a few control points). A high-level positioning chart of different representations in terms of flexibility and realism is shown in Figure 5.1.

This chapter describes a new representation that is both more realistic and flexible than previous efforts. The design space is represented as designs $\mathbf{x}$ sampled from a statistical distribution $p(\mathbf{x})$, and a generative model of this distribution is estimated using a large set of images and associated design attributes of previous designs. The key design contribution in this work is the approach of changing the product form design representation to a statistical distribution, and estimating the product form design space using large-scale data of previous designs. A methodological contribution is the use of a crowd to act as an optimization algorithm for deep generative models; namely, we crowdsource opinions on whether generated designs look realistic from varying generative models. This step is important because validation of generative models is not objective, and significant differences may exist between numerical validation metrics such as "reconstruction error" versus visual quality (Theis *et al.*, 2015).

To demonstrate these ideas, we conduct an experiment within the product form area of automobile styling. The design space distribution $p(\mathbf{x})$ is approximated using a variational autoencoder (Kingma & Welling, 2013), over a data set of 2D images and design attributes of 179,702 automobile designs from the last decade. Preliminary results show that we are able to estimate a mathematical representation that

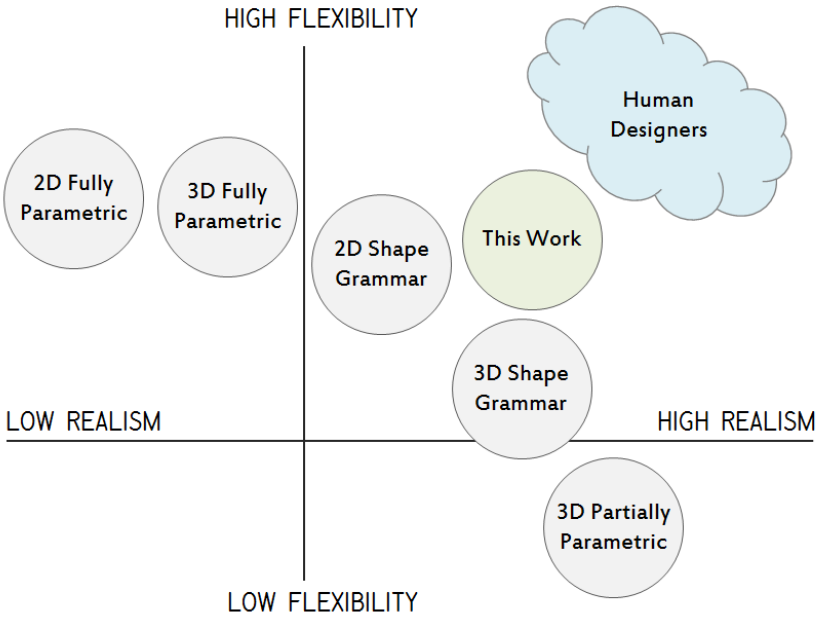


Figure 5.1: Positioning chart of product form design representations according to levels of realism and flexibility of representation.

is both realistic and flexible. We then explore this estimated design space by morphing vehicles via manupulation of design attributes such as body type, brand, and viewpoint.

This rest of this chapter is structured as follows: Section 5.2 discusses human and mathematical design representations, as well as generative models. Section 5.3 develops the mathematical representation underpinning of the conceptual design process and the deep generative model used to approximate it. Section 5.4 details the numerical and crowdsourced experiment used to estimate the design space for automobile styling. Section 5.5 explores the design space and crowdsourcing results, and discusses the implications of this design representation, limitations, and opportunities for future work. We conclude in Section 5.5.

5.2 Related Work

We discuss product form design representations in behavioral science research conducted on novice and expert designers during the conceptual design process, and in the mathematical formalization of design representations by design researchers. Next, we discuss generative models used in design and machine learning research, including the difficulty in establishing objective validation metrics for such models.

Human Designer Mental Representation

The human designer's mental representation has been studied by design researchers extensively, with much focus on behavioral differences between novice and expert designers (Dinar *et al.*, 2015; Cross, 2004b), and their mental representation of design knowledge (Chandrasegaran *et al.*, 2013b). Experts have been found to be better at representation *realism*, where realism is defined as the degree of design detail (Björklund, 2013), and ability to connect design knowledge through sketches (Kavakli & Gero, 2001b) and design analogies (Fu *et al.*, 2015; Linsey, 2007).

Expert designers have also been found to be significantly different with regards to *flexibility* during the conceptual design search process. Expert designers make smaller "leaps" between design analogies when traversing their mental design representation space (Ozkan & Dogan, 2013), and are more likely to work backwards from the design solution (Ho, 2001b), using a design problem decomposition strategy that enables "efficient" traversal of the design space.

Mathematical Design Representations

Mathematical representations of the product design space are more straightforward to compare as they are defined explicitly, thus constructing the design space according to all possible states of the representation. As noted in Section 5.1, these mathematical representations use a variety of formal elements that can be placed into

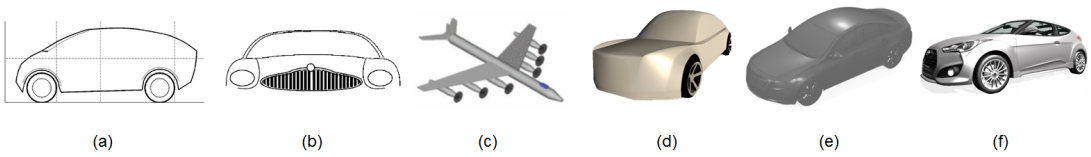


Figure 5.2: Example designs from various mathematical product form representations: (a) 2D fully parametric (Reid *et al.*, 2010a), (b) 2D shape grammar (McCormack *et al.*, 2004a), (c) 3D shape grammar (Oberhauser *et al.*, 2015), (d) 3D fully parametric (Ren *et al.*, 2013a), (e) 3D partially parametric with estimated handles (Yumer *et al.*, 2015b), (f) 3D partially parametric with hand-engineered handles (Burnap *et al.*, 2015a).

six major categories as shown in Figure 5.2. While these mathematical representations have found numerous successes, including use by real designers (Reid *et al.*, 2010a; McCormack *et al.*, 2004a; Kókai *et al.*, 2007b; Telenko *et al.*, 2016), each is limited by the tradeoff between *flexibility* and *realism* as illustrated in the positioning chart of Figure 1.

Fully parametric 2D or 3D vector representations have a high degree of flexibility since they are generally capable of morphing between all designs in the design space. This characteristic is important for the validity of results drawn from experiments using these representations. For example, assessing customer preferences using these representations enables full coverage of the space. The drawback of these representations, as shown in Figures 5.2(a) and 5.2(d), is that the resulting representations are often of limited realism due to their relatively low dimensionality.

Partially parametric 2D and 3D vector representations are manipulated with a lower dimensional set of "handles" or "control points" that affect the design representation's pixels or voxels using an attachment function. This function can be defined through a functional form, such as a kernel (Yumer *et al.*, 2015b; Murugappan *et al.*, 2013), or statistically estimated model (Yumer *et al.*, 2015a). The attachment functions often work on the entire design representation via manipulating all pixels or vertices, or by deforming the area or volume itself as shown in Figure 5.2(e). Another category shown in Figure 5.2(f) works on more fine details (e.g., headlight form and

LEDs), but requires the use of design experts to hand-engineer various parametric handles.

Opposite of the fully parametric vector representations, partially parametric representations typically are very high-dimensional (i.e., 10,000's of pixels or 100,000's of voxels), and are subsequently very realistic. This comes at the cost of limited flexibility–the extent of possible manipulations is restricted to local perturbations. For larger changes, constraints must be placed between existing designs, typically through correspondence points with very strong and perhaps unrealistic assumptions on the interpolation function (e.g., linear or quadratic) (Kókai *et al.*, 2007b).

Shape grammars, both 2D and 3D as shown in Figure 5.2(b) and Figure 5.2(c), occupy a middle ground in terms of flexibility and realism. These representations are interesting in that the design space they define is much larger than fully or partially parametric vector representation due to being combinatorial in in their composition of designs. Accordingly, while they offer flexibility across various designs, such "paths" between designs are not readily apparent. This comes with the advantage of being able to extrapolate much more reasonably as compared with vector representations. We discuss possible directions in combining random vector representations in the current work and grammar representations in Section 5.5.

Design Generative Models

Methods to generate design concepts have received attention by the design research community (Orsborn *et al.*, 2009b; Reid *et al.*, 2010a; Ren *et al.*, 2013a; Yannou *et al.*, 2008b; Petiot *et al.*, 2009) and practicing designers (Kókai *et al.*, 2007b). These methods employ the mathematical representations noted in Section 5.2 and a variety of design generation schemes.

Human-guided design selection use queries with multiple generated designs in response or as an iterative communication between human and machine. A single query

for a set of generated responses is often the focus of knowledge representation tools geared towards product form representations (Chandrasegaran *et al.*, 2013b) following early ideas from Herbert Simon (Simon, 1996). Iterative communication tools include interactive genetic algorithms (Smith, 1991; Yannou *et al.*, 2008b; Poirson *et al.*, 2013b) and more recently proposed online crowdsourcing methods (Ren *et al.*, 2013a; Ren & Papalambros, 2012a).

Deep Generative Models

Deep generative models refer to a class of hierarchical statistical models (referred to as "deep learning" in the computer science community; see (Bengio, 2009; Schmidhuber, 2015) for survey) characterized by being composed of multiple layers of nonlinear functions, with each layer connected to its adjacent layers via a set of connecting "weights." Like all generative statistical models, these deep generative models work by modeling the data distribution, using assumptions on the data space, rather than the locally connective assumptions used in graph methods. Such models have recently received renewed attention due to their successes on benchmark tasks such as 2D image object recognition (Krizhevsky *et al.*, 2012). Here we discuss three related models that form the state-of-the-art with regards to 2D image generation.

The generative adversarial network (GAN) is a generative model that has a unique parameter estimation approach (Goodfellow *et al.*, 2014). The model is divided into two parts, a generator and a discriminator; the generator is trained to generate images so that the discriminator cannot distinguish them from the ground truth images, while the discriminator is trained to discriminate generated images from the known "ground truth" images. The two parts are trained simultaneously to force the generator to produce images as similar to the ground truth images as possible, where similarity is defined by the discriminator. Experiments have shown that this model is capable of generating highly realistic images with some exceptions. In particular, since the

discriminator makes decisions based on pixel-level distance metrics, the generator can make unrealistic mistakes important to human reviewers, e.g., a face with a displaced nose.

The deconvolutional neural network is a multi-layer model composed of fully connected layers followed by two sets of convolutional layers–one tasked to generate design images and the other to generate segmentation masks of the design (Dosovitskiy *et al.*, 2014). This model takes multiple design attribute to be generated (e.g., types of chairs). The model assembles a deterministic function that maps a set of input attributes to one output; however, this modelling assumption does not align well with our goal of capturing uncertainty from design attributes–we discuss this in detail in Section 5.3.

The variational autoencoder (VAE) (Kingma & Welling, 2013) used in this work is an advanced version of the deconvolutional neural network, with major differences in the method of statistically estimating the model parameters of the model in its parametrization to introduce randomness to the generating process. A detailed introduction of the VAE model is in Section 5.3.

Validation of Generative Models

One challenging issue inherent to generative models is their the lack of straightforward validation. The requirements of this validation are twofold: The model needs to generate realistic 2D images that can be recognized by human viewers as a particular category of objects (e.g., cars), while these images must be different from any image the model has seen in the data set, or otherwise overfitting on the data set would model a simple solution of memorizing known training images. While the former requirement leads the model to produce similar images to the ones used in training, the later one forces the model in the opposite direction (i.e., generalization via interpolation and extrapolation).

Consequently, it is nontrivial to establish a quantitative validation that reflects model performance on both requirements. Research has been done using measurements such as pixel-level Euclidean distance, image retrieval from the known data set, and structured similarity indices from nearest neighbor images in the training set. However, none of these methods can give direct validation regarding the two requirements. In many cases, a generated 2D image that results in a favorable score under numerical measures scores very low on visual quality as assessed by a human (Theis *et al.*, 2015). To address this issue, we propose to utilize a crowd as a direct validation of the model's capability to generate realistic images.

5.3 Problem Formulation

The problem formulation begins with assuming a fictitious conceptual design scenario involving three ingredients: (1) A "true" product form design space $\mathcal{X}$ containing the product forms of all 2D images capturing what it means to be the particular design (e.g., a passenger vehicle); (2) a "complete" (possibly infinite) list of design attributes, denoted $\mathbf{a}^*$, and obtained by being the exact set of design attributes most preferred by the targeted customer; and (3) a "perfect" design tool f^*, able to map deterministically a single complete design attribute list $\mathbf{a}^*$ to a single design $\mathbf{x} \in \mathcal{X}$:

$$\mathbf{x} = f^*(\mathbf{a}^*) \tag{5.1}$$

In reality, we do not have access to this complete set of design attributes $\mathbf{a}^*$ (e.g., the customer would most prefer exactly these bodyline curves, taillight shape and illumination, etc.), and must instead settle for a dramatically smaller finite set of design attributes $\mathbf{a}$ (e.g., the customer would prefer a 'Cadillac' 'Coupe'). We now have a massive number of unknown latent variables called "design features" denoted $\mathbf{h}$ as introduced in (Burnap *et al.*, 2016). In other words, the previous complete list of

design attributes is now partitioned into known design attributes and unknown design features $\mathbf{a}^* = \{\mathbf{a}, \mathbf{h}\}$. This introduces uncertainty into our originally deterministic function, which may now be represented according to a statistical distribution p_* with unknown functional form:

$$\mathbf{h} \sim p_*(\mathbf{h}). \tag{5.2}$$

Since we do not know this functional form, we instead assume the uncertainty from the random vector $\mathbf{h}$ may be captured by a distribution parametrized by θ, giving us the conditional distribution we aim to estimate:

$$\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{a}) \tag{5.3}$$

Practically estimating the parameters of this distribution is challenging due to the high dimensionality of $\mathcal{X}$, which itself is a subset of a universal domain (Gero & Maher, 2013) of all 256-bit RGB values of the number of pixels forming the 2D image (i.e., $N = 256^{(3 \cdot \text{pixels})}$). Accordingly, we turn to a variational approximation (Wainwright & Jordan, 2008) of the conditional likelihood of the data, one that will be particularly suited to online mode-seeking optimization methods.

This requires introduction of a latent random vector $\mathbf{z}$ as a tool to make the variational approximation using a tractable distribution q_ϕ:

$$\begin{aligned} \log p_\theta(\mathbf{x}|\mathbf{a}) &= \sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{a}) \log p_\theta(\mathbf{x}|\mathbf{a}) \\ &= -\sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{a}) \log p_\theta(\mathbf{z}|\mathbf{x}, \mathbf{a})) \\ &+ \sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{a})) \log p_\theta(\mathbf{x}, \mathbf{z}|\mathbf{a})) \\ &= KL\left(q_{\mathbf{z}|\mathbf{x},\mathbf{a}} || p_{\mathbf{z}|\mathbf{x},\mathbf{a}}\right) + \mathcal{L}\left(\theta, \phi; \mathbf{x}\right) \end{aligned} \tag{5.4}$$

where $KL\left(q_{\mathbf{z}|\mathbf{x},\mathbf{a}}||p_{\mathbf{z}|\mathbf{x},\mathbf{a}}\right)$ is the KL divergence, which is always non-negative. Therefore, the second term $\mathcal{L}(\theta,\phi;\mathbf{x})$ becomes a lower bound of the conditional likelihood given in Equation (5.3), and becomes the objective function we seek to maximize.

We expand $\mathcal{L}(\theta,\phi;\mathbf{x})$ into three terms, which are then amenable to the deep generative model in Section 5.3.1:

$$\mathcal{L}(\theta,\phi;\mathbf{x}) = \sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x},\mathbf{a})\left(\log p_\theta(\mathbf{x},\mathbf{z}|\mathbf{a}) - \log q_\phi(\mathbf{z}|\mathbf{x},\mathbf{a})\right) \tag{5.5}$$
$$= \sum_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x},\mathbf{a})\left(\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{a}) + \log p_\theta(\mathbf{z}|\mathbf{a}) - \log q_\phi(\mathbf{z}|\mathbf{x},\mathbf{a})\right).$$

5.3.1 Deep Generative Model

We estimate the parameters θ and ϕ for the conditional likelihood given in Equation (5.5) using a hierarchical parametric model (i.e., "deep learning") that exploits invariance in 2D images, as well as optimization techniques to obtain point estimates to the values of these parameters. In particular, we use a variational autoencoder (VAE), a variational Bayesian approach introduced by Kingma and Welling (Kingma & Welling, 2013) that learns a directed probabilistic model by approximating the posterior expectation with a reparametrization trick.

The VAE is made up of two networks: an "encoder" that transforms the 2D images within the data space to a latent representation (i.e., the last term in Equation (5.5), and a generative "decoder" model that transforms the latent representation back to a 2D image reconstruction in the data space, i.e., the first term in Equation (5.5). We use an extension to the VAE that includes conditioning on additional data (Kingma *et al.*, 2014; Yan *et al.*, 2015; Louizos *et al.*, 2015), in our case known design attributes $\mathbf{a}$ associated with a given design $\mathbf{x}$. These conditional terms allow the latent representation to instead focus on encoding the uncertainty from features $\mathbf{h}$ not contained in the known design attributes $\mathbf{a}$:

$$\mathbf{z} \sim \text{Encoder}(\mathbf{x}) = q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{a}) \tag{5.6}$$

$$\hat{\mathbf{x}} \sim \text{Decoder}(\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{a}) \tag{5.7}$$

The reparametrization trick discussed (Kingma & Welling, 2013) expresses our introduced latent random variable $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{a})$ with a deterministic variable $\mathbf{z} = g_\phi(\epsilon, \mathbf{x}, \mathbf{a})$, where ϵ is an independent "auxiliary" random variable, and $g_\phi(\cdot)$ is some vector-valued function parametrized by ϕ. Further, we approximate this axillary variable using Monte Carlo sampling:

$$\begin{aligned} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{a})}\left[f\left(\mathbf{z}\right)\right] &= \mathbb{E}_{p(\epsilon)}\left[f\left(g_\phi\left(\epsilon, \mathbf{x}, \mathbf{a}\right)\right)\right] \\ &\approx \frac{1}{L}\sum_{l=1}^{L} f\left(g_\phi\left(\epsilon^{(l)}, \mathbf{x}, \mathbf{a}\right)\right) \text{ with } \epsilon^{(l)} \sim p\left(\epsilon\right) \end{aligned} \tag{5.8}$$

in which l denotes Monte Carlo draws and L denotes the total number of draws. Using Equation (5.8), we reparametrize the lower bound of the conditional likelihood we are after in Equation (5.5):

$$\begin{aligned} \mathcal{L}\left(\theta, \phi; \mathbf{x}\right) \quad \approx \quad & \frac{1}{L}\sum_{l=1}^{L} \log p_\theta\left(\mathbf{x}|\mathbf{z}^{(l)}, \mathbf{a}\right) \\ & + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{z}^{(l)}|\mathbf{a}) \\ & - \frac{1}{L}\sum_{l=1}^{L} q_\phi\left(\mathbf{z}^{(l)}|\mathbf{x}, \mathbf{a}\right) \\ \text{where} \quad & \mathbf{z}^{(l)} = g_\phi\left(\epsilon^{(l)}, \mathbf{x}\right), \epsilon^{(l)} \sim p\left(\epsilon\right) \end{aligned} \tag{5.9}$$

Lastly, we define $q_\phi(\mathbf{z}|\mathbf{x}), p_\theta(\mathbf{z}), p_\theta(\mathbf{x}|\mathbf{z})$ as Gaussian distributions, whose parameters θ and ϕ we estimate using the VAE:

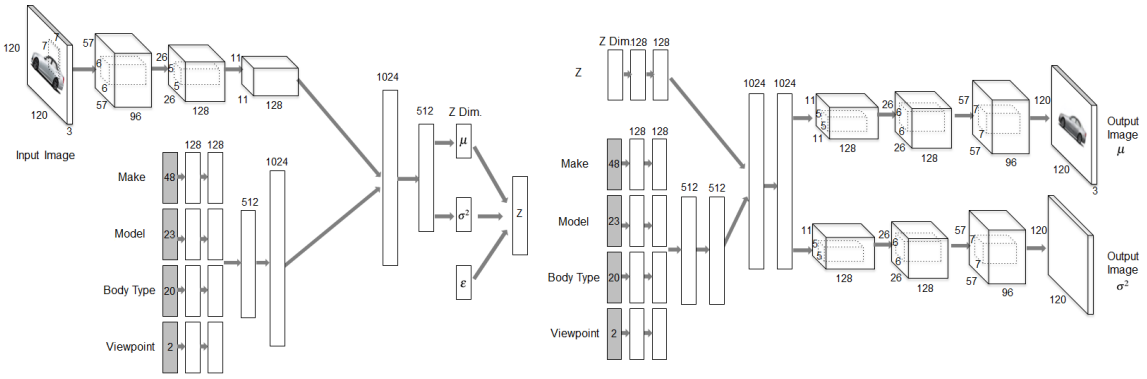


Figure 5.3: Deep generative model architecture of variational autoencoder; on the left is the encoder, while the right is the decoder. Shaded boxes represent inputs, white boxes represent fully connected layers, and rectangular prisms represent convolutional and pooling layers in the encoder, and upsampling layers in the decoder.

$$q_\phi(\mathbf{z}|\mathbf{x},\mathbf{a}) = \mathcal{N}\left(\mathbf{z};\mu_\phi(\mathbf{x},\mathbf{a}),\sigma^z_\phi(\mathbf{x},\mathbf{a})\right) \tag{5.10}$$

$$p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I}) \tag{5.11}$$

$$p_\theta(\mathbf{x}|\mathbf{z},\mathbf{a}) = \mathcal{N}\left(\mathbf{x};\mu(\mathbf{z},\mathbf{a}),\sigma^2(\mathbf{z},\mathbf{a})\right) \tag{5.12}$$

5.3.2 Model Architecture

The architecture of a deep hierarchical model concerns the types of (i) "layers", i.e., vectors of functions that define function compositions between layers; (ii) "neurons" or "filters" making up the various layers, particularly their functional form; and (iii) connectivity linking layers to each other via parameters θ and ϕ.

The chosen architecture significantly influences the performance of the deep generative model, as architecture decisions constrain the flow of information throughout the model. Poor architecture choices increase the number of parameters of the model θ and ϕ or sub-optimal generative performance. For a VAE, a number of special lay-

ers is used to reduce the number of parameters while trading off information capture of the underlying data distribution. We show in Figure 5.3.1 the model architecture that uses four types of layers as described below:

Fully Connected Layers

With a fully connected layer, the input $\mathbf{x} \in \mathbb{R}^{BxN}$ and the output $\mathbf{y} \in \mathbb{R}^{BxM}$ are associated with the function of $\mathbf{y} = f(\mathbf{x}^T\mathbf{w} + \mathbf{b})$, where $\mathbf{w} \in \mathbb{R}^{NxM}, \mathbf{b} \in \mathbb{R}^M$, and $f(\cdot)$ denotes a nonlinear function–in our case, the Rectified Linear (ReLu).

Convolutional Layers

Convolutional layers capture the notion that there are translation and rotation invariance, such as local regions forming image components that exist globally across the image. Such convolutional filters greatly reduce the number of parameters necessary in the layer relative to a fully connected layer, while still capturing a similar amount of information.

Similar to the fully connected layer, in a convolutional layer the input $\mathbf{x}$ and the output $\mathbf{y}$ are associated with the function of $\mathbf{y} = f(\mathbf{w} \bigotimes \mathbf{x} + \mathbf{b})$, where $\bigotimes$ denotes the 2D convolution operation and $f(\cdot)$ denotes a nonlinear function, in which we use Rectified Linear (ReLu), except the last layer where output is produced in which no nonlinearity function is employed.

Pooling Layers

Because pixel value information is highly redundant in images (i.e., neighbor pixels values are highly correlated) additional measurements are taken to reduce the number of parameters in the model. In a pooling layer, one output value will be used to replace a square area of input values. In this work, we use max pooling layers with a pooling size of 2 by 2, i.e., the maximum value of the 2 by 2 pixels in an area will be used,

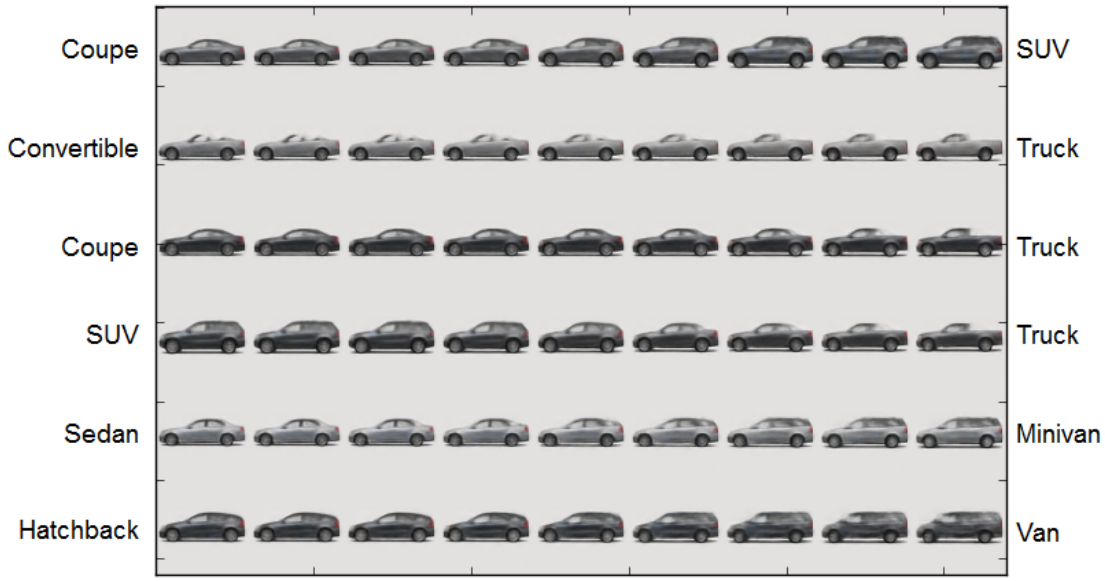


Figure 5.4: Morphing between various body types from the estimated product form design space.

and the rest are discarded thus reduced the parameters by a factor of 2 for each dimension.

Upsampling Layers

Upsampling refers to the inverse of a "pooling" operation, which is used to "upsample" the coded information back to the same dimension as the input images. This operation necessarily loses information; however, the choices for such approximate inversion are varied (e.g., fixed location upsampling, average upsampling, and upsampling with switch units). In the current work, we use average upsampling.

5.4 Experiment: Generating the Last Decade of Automobiles

Our goal in the experiment was to statistically estimate the design space $p_\theta(\mathbf{x})$ to obtain a mathematical representation with realism and flexibility advantages as described in Section 5.1, using the model described in Section 5.3.1 and optimized

using both numerical techniques and crowdsourcing.

Dataset

The data set consisted of 179,702 data points, with each point made up of a 2D image and four design attributes–make, model, body type, and viewpoint–with corresponding dimensionality shown in Figure 5.3.1 . Each 2D image was downscaled to 120x120 pixels using OpenCV, an open source computer vision library (Bradski & Kaehler, 2008). We then split this data of previous designs into a "training set" and "validation set" with a 3:1 split ratio.

Numerical Parameter Estimation

The variational autoencoder described in Section 5.3.1 requires a number of hyperparameters (i.e., user-defined values such as learning rate and batch size) during the statistical estimation of the parameter sets θ and ϕ, as well as hyperparameters of the architecture itself (e.g., number of neurons or filter in a layer). We give these architecture hyperparameters in Figure 5.3.1. This architecture was implemented using Theano (Bergstra *et al.*, 2010), an open source symbolic differentiation library with a graph-based compiler and GPU-acceleration support.

First-order methods have shown to be often better suited to estimation of deep generative models, particularly when extended with terms that mitigate being affected by saddle points (Dauphin *et al.*, 2014) and sharp discontinuities (Szegedy *et al.*, 2013). For these reasons, we use the ADAM optimizer (Kingma & Adam, 2015), which has is particularly suited to parameter estimation of deep generative models. We use the ADAM optimization parameter of $\beta_1 = 0.1$ and $\beta_2 = 0.2$ with a learning rate of $\alpha = 0.0002$. Moreover, estimation of the parameter sets θ and ϕ in practice requires the use of "mini-batches" due to large data set sizes; accordingly, we used a mini-batch size of 100.

Crowdsourced Hyperparameter Estimation

The goal for crowdsourced hyperparameter estimation was to assess whether there were significant differences in visual quality of generated 2D images, when varying the number of latent random variables Z used in the model architecture as described in Section 5.3.1. This assessment was performed as it has been shown that using numerical performance measures (e.g., log-likelihood) does not necessarily correspond with human perception of visual quality (Theis *et al.*, 2015). While certain theories (e.g., manifold hypothesis (Bengio *et al.*, 2015)) suggest that the effective dimensionality may not be best modeled as fixed, the addition of humans-in-the-loop may provide complementary advantages.

5.4.0.1 Participants

A total of 69 participants were gathered using the crowdsourcing platform Amazon Mechanical Turk using an open call and a monetary incentive. We filtered out participants that "clicked through" the online application if their responses took less than 3 seconds per "survey question."

5.4.0.2 Procedure

A web application with a database backend was developed to collect participant responses to generated 2D images with varying model architecture hyperparameters. Participants were first directed to a home page, which described the instructions for inputting visual quality responses to 2D images.

After clicking to proceed past the instructions, participants were presented with an ordered set of 2D images, and asked to select the 2D image that was most realistic. Each ordered set contained three 2D images, corresponding to three settings of the hyperparameters controlling the number of dimensions (i.e., 32, 128, 256 dimensions) in the latent representation Z. Each ordering was random, in order to not bias

Figure 5.5: Morphing between various brands from the estimated product form design space.

participants, while all the same design attributes were held constant (e.g., 'Cadillac,' 'coupe.'). The possible set of 2D image triplets contained all pairwise combinations of bodytypes from the sideview of the vehicle.

Participants were only allowed to click one of the three 2D images, and were able to change their selection. After participants completed 20 randomly selected 2D image triplets, they were redirected to web page thanking them for their time and presenting them with a unique code for monetary redemption.

5.5 Results and Discussion

We explore the estimated product form design space by morphing between various pairs of design attributes. To show the flexibility of the estimated mathematical representations $p_\theta(\mathbf{x})$, we morph between various body types in Figure 5.4, various brands in Figure 5.5, and rotational viewpoints in Figure 5.6.

The designs $\mathbf{x}$ sampled for these results are all artificially generated (i.e., none are in the dataset). Moreover, we show multiple steps in between each design attribute

Figure 5.6: Rotations of various body types from the estimated product form design space.

pair to indicate that we are not overfitting on the data set, as none of these generated designs exist. In particular, we observe that the visual quality of the generated designs is uniform across various morphing steps between known design attributes (e.g., from coupe to SUV); this reinforces the notion that we are not simply overfitting, and instead we are estimating the true product design space $p_\theta(\mathbf{x}|\mathbf{a})$.

The motivation for this work was developed in part from working with practicing designers in the automotive industry and recognizing the necessity of a design representation that can morph between various brands and body types, yet realistic enough to convey sufficient meaning to designers (Burnap *et al.*, 2015a). This representation is not limited to brand studies. A number of design questions can be explored. For example, we show in Figure 5.7 a generated vehicle between a 'sedan' and an 'SUV,' which is currently the fastest growing design segment in the automotive market. This type of design generation can serve as inspiration to designers working on designs for new market segments (Hartley, 1996b).

Figure 5.8 shows the results of the crowdsourced parameter optimization. Preliminary results indicate that we cannot conclusively state whether the crowd was able

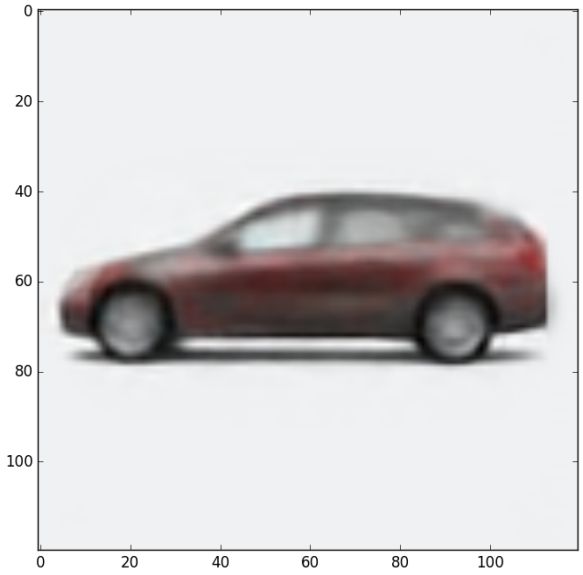


Figure 5.7: Generated vehicle between 'sedan' and 'SUV' for randomly set brand that looks like a 'crossover.'

to discriminate between various hyperparameter settings during the design space estimation. Further research is required into using a crowd to fine-tune parameters affecting image quality after an initial computer-only optimization is performed.

Thus, the hypothesized value of using crowdsourced optimization requires deeper investigation. If the crowd is shown to improve the statistical estimation procedure, we may be able to build more robust crowd-powered optimization systems for these generative models. The current approach is not in real time; however, a worthwhile goal may be to build a real-time optimization loop including the crowd, particularly if incentivized as in the emerging area of gamification (Ren *et al.*, 2015b).

Limitations and Future Work

Interpretation of the latent space remains a challenging and potentially rewarding goal in thus research. Nonlinear predictive models, particularly those recently popu-

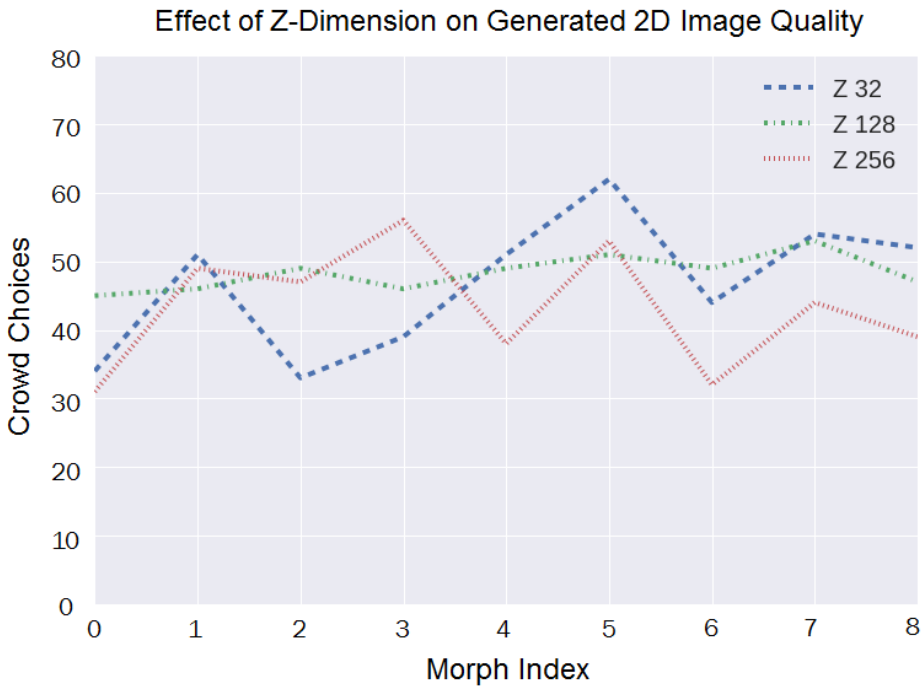


Figure 5.8: Effect of various number of **z** random variables in hidden representation on generated 2D image quality as assessed by crowd. Morph index refers to how far between two known design attributes a design was morphed– e.g., 0 and 8 may be 'convertible' and 'truck', respectively.

larized by data science studies around large-scale datasets, offer significantly improved generalization performance, and thus significantly improved capture of the underlying physics of the design task, e.g., preference prediction, design space representation, and market segmentation.

These models are in contrast with the interpretable linear models commonly used in design task modeling, which often work on strong modeling assumptions comprised of main effects and pairwise interaction terms, and thus neglect all other statistical dependencies amongst design variables. Future work towards interpretation of these latent representation may offer much deeper insight into underlying design preferences, translated into actionable design decisions that capture how the designer can adjust design attributes to elicit desired preferences within a specific population.

We show in Figure 5.9 an example of the design feature "color" that we do not currently capture. While this feature may be simpler to capture using crowds, numer-

Figure 5.9: Generated design displaying the design features 'color,' which we do not yet control.

ous other features still exist that are not as straightforward. Current work towards such "feature interpretation" has shown preliminary promise, including data-driven approaches to predict which visual features of a design most elicit attention (Pan *et al.*, 2016). These approaches aim to move into the causality behind features in deep convolutional neural networks (Zeiler & Fergus, 2014; Simonyan *et al.*, 2013).

Interpreting such features may lead to new shape grammars. The combination of top-down statistical estimation of the design space and bottom-up definition of the space using shape grammars may be a valuable direction for future research. Validation such combination may be aided by methods that use humans-in-the-loop, such as online crowdsourcing (Burnap *et al.*, 2015c), or in-person eye tracking (Reid *et al.*, 2013b; Du & MacDonald, 2014).

5.6 Summary

Human designers use a mental design representation of product form that is both flexible and realistic, allowing efficient exploration of the design space during the conceptual design process. Mathematical representations of the product form design space impose constuctivist restrictions on the design space and trade off representation flexibility for representation realism.

We changed a statistical distribution as a mathematical representation that is more flexible and realistic than previously proposed representations. We formulated this representation by assuming a "perfect" conceptual design scenario and progressively introduced uncertainty as dictated by the real world. We approximated this true statistical distribution using a deep generative model, in particular a variational autoencoder, which is amenable to efficient computing of large-scale data sets.

We conducted an experiment in the product form domain of automotive styling, using design attributes and 2D images of automobile designs from the last decade. The results showed that we are able to find a design representation that is both flexible and realistic in exploring the design space over design attributes such as body type, brand, and viewpoint. We also examined using a crowd to improve the parameter estimation process of the deep generative model; our preliminary results showed that we are not yet able to improve our generative model results to statistically significant levels in this way.

Lastly, we discussed a number of possible improvements to this work within the emerging area of data-driven design. In particular, interpretation of design features otherwise wrapped up in uncertainty offers design researchers and practicing designers opportunities for valuable design insight. Further investigation into crowdsourcing mechanisms, real-time and gamified, may prove fruitful. Aligning design augmentation tools with practicing designers and design researchers who study expert designers remains important and can further the value of design automation tools.

CHAPTER VI

Why does Crowdsourcing Fail for Subjective Preferences?

6.1 Context: Which passenger vehicle would you purchase?

Much research has been devoted to develop design preference models that predict customer design choices. A common approach is to: (i) Collect a large database of previous purchases that includes customer data, e.g., age, gender, income, and purchased product design data, e.g., # cylinders, length, curb weight — for an automobile; and (ii) statistically infer a design preference model that links customer and product variables, using conjoint analysis or discrete choice analysis such as logit, mixed logit, and nested logit models (**?**McFadden & Train, 2000).

However, a customer may not purchase a vehicle solely due to interactions between these two sets of variables, e.g., a 50-year old male prefers 6-cylinder engines. Instead, a customer may purchase a product for more 'meaningful' design attributes that are functions of the original variables, such as environmental sustainability or sportiness (**?**Norman, 2007). These meaningful intermediate functions of the original variables, both of the customer and of the design, are hereafter termed *features*. We posit that using customer and product features, instead of just the original customer and product variables, may increase the prediction accuracy of the design preference model.

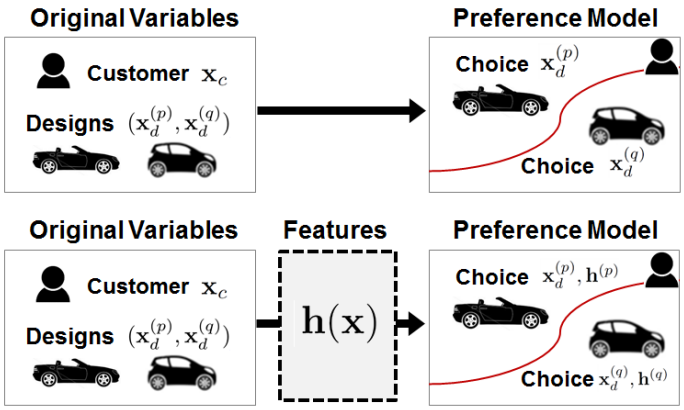


Figure 6.1: The concept of feature learning as an intermediate mapping between variables and a preference model. The diagram on top depicts conventional design preference modeling (e.g., conjoint analysis) where an inferred preference model discriminates between alternative design choices for a given customer. The diagram on bottom depicts the use of features as an intermediate modeling task.

Our goal then is to find features that improve this preference prediction accuracy. To this end, one common approach is to ask design and marketing domain experts to choose these features intuitively, such as a design's social context (**?**) and visual design interactions (**?**). For example, eco-friendly vehicles may be a function of miles per gallon (MPG) and emissions, whereas environmentally active customers may be a function of age, income, and geographic region. An alternative explored in this paper is to find features 'automatically' using feature learning methods studied in computer science and statistics. As shown in Figure 6.1, feature learning methods create an intermediate step between the original data and the design preference model by forming a more efficient "feature representation" of the original data. Certain well-known methods such as principal component analysis may be viewed similarly, but more recent feature learning methods have shown impressive results in 1D waveform prediction (**?**) and 2D image object recognition (**?**).

We conduct an experiment on automobile purchasing preferences to assess whether three feature learning methods increase design preference prediction accuracy: (1) principal component analysis, (2) low-rank + sparse matrix decomposition, and (3)

exponential family sparse restricted Boltzmann machines (**?**Salakhutdinov *et al.*, 2007). We cast preference prediction as a binary classification task by asking the question, "given customer $\mathbf{x}$, do they purchase vehicle p or vehicle q." Our data set is comprised of 1,161,056 data points generated from 5582 real passenger vehicle purchases in the United States during model year 2006 (MY2006).

The first contribution of this work is an increase of preference prediction accuracy by 2%-7% just using simple "single-layer" feature learning methods, as compared with the original data representation. These results suggest features indeed better represent the customer's underlying design preferences, thus offering deeper insight to inform decisions during the design process. Moreover, this finding is complementary to recent work in crowdsourced data gathering (**?**Panchal, 2015a) and nonlinear preference modeling (**?**Evgeniou *et al.*, 2007a)) since they do not affect the preference model or data set itself.

The second contribution of this work is to show how features may be used in the design process. We show that feature interpretation and feature visualization offer designers additional tools for augmenting design decisions. First, we interpret the most influential pairings of vehicle features and customer features to the preference task, and contrast this with the same analysis using the original variable representation. Second, we visualize the theoretically optimal vehicle for a given customer within the learned feature representation, and show how this optimal vehicle, which does not exist, may be used to suggest design improvements upon current models of vehicles that do exist in the market.

Methodological contributions include being the first to use recent feature learning methods on heterogeneous design and marketing data. Recent feature learning research has focused on homogeneous data, in which all variables are real-valued numbers such as pixel values for image recognition (**?**Lee *et al.*, 2011); in contrast, we explicitly model the heterogeneous distribution of the input variables, for example

'age' being a real-valued variable and 'General Motors' being a categorical variable. Subsequently, we give a number of theoretical extensions: First, we use exponential family generalizations for the sparse restricted Boltzmann machines, enabling explicit modeling of statistical distributions for heterogeneous data. Second, we derive theoretical bounds on the reconstruction error of the low-rank + sparse matrix decomposition feature learning method.

This paper is structured as follows: Section 6.2 discusses efforts to increase prediction accuracy by the design community, as well as feature learning advances in the machine learning community. Section 6.3 sets up the preference prediction task as a binary classification problem. Section 6.4 details three feature learning methods and their extension to suit heterogeneous design and market data. Section 6.5 details the experimental setup of the preference prediction task, followed by results showing improvement of preference prediction accuracy. Section 6.7 details how features may be used to inform design decisions through feature interpreation and feature visualization. Section 6.8 concludes this work.

6.2 Related Work

Design preference modeling has been investigated in design for market systems, where quantitative engineering and marketing models are linked to improve enterprise-wide decision making (**?**Lewis *et al.*, 2006; Michalek *et al.*, 2005). In such frameworks, the design preference model is used to aggregate input across multiple stakeholders, with special importance on the eventual customer within the targeted market segment (**?**).

These design preference models have been shown to be especially useful for the design of passenger vehicles, as demonstrated across a variety of applications such as engine design (**?**), vehicle packaging (**?**), brand recognition (**?**), and vehicle styling (**?**Reid *et al.*, 2012; Sylcott *et al.*, 2013a). Connecting many of these research efforts is

the desire for improved prediction accuracy of the underlying design preference model. With increased prediction accuracy, measured using "held out" portions of the data, greater confidence may be placed in the fidelity of the resulting design conclusions.

Efforts to improve prediction accuracy involve: (i) Developing more complex statistical models to capture the heterogeneous and stochastic nature of customer preferences; examples yuinclude mixed and nested logit models (**?**Berkovec & Rust, 1985), consideration sets (**?**), and kernel-based methods (**?**Evgeniou *et al.*, 2007a; Ren *et al.*, 2013b); and (ii) creating adaptive questionnaires to obtain stated information more efficiently using a variety of active learning methods (**?**Abernethy *et al.*, 2008).

This work is different from (i) above in that the set of features learned is agnostic of the particular preference model used. One can just as easily switch out the l^2 logit design preference model used in this paper for another model, whether it be mixed logit or a kernel machine. This work is also different from (ii) in that we are working with a set of revealed data on actual vehicle purchases, rather than eliciting this data through a survey. Accordingly, this work is among recent efforts towards data-driven approaches in design (**?**), including design analytics (**?**) and design informatics (**?**), in that we are directly using data to augment existing modeling techniques and ultimately suggest actionable design decisions.

Feature learning

Feature learning methods capture statistical dependencies implicit in the original variables by "encoding" the original variables in a new feature representation. This representation keeps the number of data the same while changing the length of each data point from M variables to K features. The idea is to minimize an objective function defining the reconstruction error between the original variables and their new feature representation. If this representation is more meaningful for the discriminative design preference prediction task, we can use the same supervised model (e.g., logit

model) as before to achieve higher predictive performance. More details are given in Section 6.4.

The first feature learning method we examined is principal component analysis (PCA). While not conventionally referred to as a feature learning method, PCA is chosen for its ubiquitous use and its qualitative difference from the other two methods. In particular, PCA makes the strong assumption that the data is Gaussian noise distributed around a linear subspace of the original variables, with the goal of learning the eigenvectors spanning this subspace (**?**). The features in our case are the coefficients of the original variables when projected onto this subspace or, equivalently, the inner product with the learned eigenvectors.

The second feature learning method is low-rank + sparse matrix decomposition (LSD). This method is chosen as it defines the features implicitly withing the preference model. In particular, LSD decomposes the "part-worth" coefficients contained in the design preference model (e.g., conjoint analysis or discrete choice analysis) into a low-rank matrix plus a sparse matrix. This additive decomposition is motivated by results from the marketing literature suggesting certain purchase consideration are linearly additive (**?**), and thus well captured by decomposed matrices (**?**). An additional motivation for a linear decomposition model is the desire for interpretability (**?**). Predictive consumer marketing oftentimes uses these learned coefficients to work hand-in-hand with engineering design to generate competitive products or services (**?**). Such advantages are bolstered by separation of factors captured by matrix decomposition, as separation may lead to better capture of heterogeneity among market segments (**?**). Readers are referred to (**?**) for further in-depth discussion.

The third feature learning method is the exponential family sparse restricted Boltzmann machine (RBM) (**?**Lee *et al.*, 2008). This method is chosen as it explicitly represents the features, in contrast with the LSD. The method is a special case of a Boltzmann machine, an undirected graph model in which the energy associated within

an energy state space defines the probability of finding the system in that state (**?**). In the RBM, each state is determined by both visible and hidden nodes, where each node corresponds to a random variable. The visible nodes are the original variables, while the hidden nodes are the feature representation. The "restricted" portion of the RBM refers to the restriction on visible-visible connections and hidden-hidden connections, later detailed and depicted in in Section 6.4 and Figure 6.4, respectively.

All three feature learning methods are considered "simple" in that they are single-layer models. The aforementioned results in 1D waveform speech recognition and 2D image object recognition have been achieved using hierarchical models, built by stacking multiple single-layer models. We chose single-layer feature learning methods here as an initial effort and to explore parameter settings more easily; as earlier noted, there is limited work on feature learning methods for heterogeneous data (e.g., categorical variables) and most advances are currently only on homogeneous data (e.g., real-valued 2D image pixels).

6.3 Preference Prediction as Binary Classification

We cast the task of predicting a customer's design preferences as a binary classification problem: Given customer j, represented by a vector of heterogeneous customer variables $\mathbf{x}_c^{(j)}$, as well as two passenger vehicle designs p and q, each represented by a vector of heterogeneous vehicle design variables $\mathbf{x}_d^{(p)}$ and $\mathbf{x}_d^{(q)}$, which passenger vehicle will the customer purchase? We use a real data set of customers and their passenger vehicle purchase decisions as detailed below (**?**).

Customer and vehicle purchase data from 2006

The data used in this work combines the Maritz vehicle purchase survey from 2006 (**?**), the Chrome vehicle variable database (**?**), and the 2006 estimated U.S. state income and living cost data from the U.S. Census Bureau (**?**) to create a data

Table 6.1: Customer variables $\mathbf{x}_c$ and their variable types

Customer Variable	Type	Customer Variable	Type
Age	Real	U.S. State Cost of Living	Real
Number of House Members	Real	Gender	Binary
Number of Small Children	Real	Income Bracket	Categorical
Number of Med. Children	Real	House Region	Categorical
Number of Large Children	Real	Education Level	Categorical
Number of Children	Real	U.S. State	Categorical
U.S. State Average Income	Real		

set with both customer and passenger vehicle variables. These combined data result in a matrix of purchase records, with each row corresponding to a separate customer and purchased vehicle pair, and each column corresponding to a variable describing the customer (e.g., age, gender, income) or the purchased vehicle (e.g., # cylinders, length, curbweight).

From this original data set, we focus only on the customer group who bought passenger vehicles of size classes between mini-compact and large vehicles, thus excluding data for station wagons, trucks, minivans, and utility vehicles. In addition, purchase data for customers who did not consider other vehicles before their purchases were removed, as well data for customers who purchased vehicles for another party.

The resulting database contained 209 unique passenger vehicle models bought by 5582 unique customers. The full list of customer variables and passenger vehicle variables can be found in Tables 6.1 and 6.2. The variables in these tables are grouped into three unit types: Real, binary, and categorical, based on the nature of the variables.

Table 6.2: Design variables $\mathbf{x}_d$ and their variable types

Design Variable	Type	Design Variable	Type
Invoice	Real	AWD/4WD	Binary
MSRP	Real	Automatic Transmission	Binary
Curbweight	Real	Turbocharger	Binary
Horsepower	Real	Supercharger	Binary
MPG (Combined)	Real	Hybrid	Binary
Length	Real	Luxury	Binary
Width	Real	Vehicle Class	Categorical
Height	Real	Manufacturer	Categorical
Wheelbase	Real	Passenger Capacity	Categorical
Final Drive	Real	Engine Size	Categorical
Diesel	Binary		

Choice set training, validation, and testing split

We converted the data set of 5582 passenger vehicle purchases into a binary choice set by generating all pairwise comparisons between the purchased vehicle and the other 208 vehicles in the data set for all 5582 customers. This resulted in $N = 1,161,056$ data points, where each datum indexed by n consisted of a triplet (j, p, q) of a customer indexed by j and two passenger vehicles indexed by p and q, as well as a corresponding indicator variable $y^{(n)} \in \{0, 1\}$ describing which of the two vehicles was purchased.

This full data were then randomly shuffled, and split into training, validation, and testing sets. As previous studies have shown the impact on prediction performance given different generations of choice sets (**?**), we created 10 random shufflings and subsequent data splits of our data set, and run the design preference prediction experimental procedure of Section 6.5 on each one independently. This work is therefore complementary to studies on developing appropriate choice set generation schemes such as (**?**). Full details into the data processing procedure are given in Section 6.5.

Bilinear design preference utility

We adopt the conventions of utility theory for the measure of customer preference over a given product (**?**). Formally, each data point consists of a pairwise comparison between vehicles p and q for customer j , with corresponding customer variables $\mathbf{x}_c^{(j)}$ for $j \in \{1, \ldots, 5582\}$ and original variables of the two vehicle designs, $\mathbf{x}_d^{(p)}$ and $\mathbf{x}_d^{(q)}$ for $p, q \in \{1, \ldots, 209\}$. We assume a bilinear utility model for customer j and vehicle p:

$$U_{jp} = \left[\text{vec} \left(\mathbf{x}_c^{(j)} \otimes \mathbf{x}_d^{(p)} \right)^T, \left(\mathbf{x}_d^{(p)} \right)^T \right] \omega, \tag{6.1}$$

where $\otimes$ is an outer product for vectors, $\text{vec}\,(\cdot)$ is vectorization of a matrix, $[\cdot, \cdot]$ is concatenation of vectors, and ω is the part-worth vector.

Design preference model

The preference model refers to the assumed relationship between the bilinear utility model described in Section 6.3 and a label indicating which of the two vehicles the customer actually purchased. While the choice of preference model is not the focus of this paper, we pilot-tested popularly used models including l^1 and l^2 logit model, naïve Bayes, l^1 and l^2 linear as well as kernelized support vector machine, and random forests.

Based on these pilot results, we chose the l^2 logit model due to its widespread use in the design and marketing communities (**?**Fuge, 2015); in particular, we used the primal form of the logit model. Equation (6.2) captures how the logit model describes the probabilistic relationship between customer j's preference for either vehicle p or vehicle q as a function of their associated utilities given by Equation (6.1). Note that ϵ are Gumbel-distributed random variables accounting for noise over the underlying utility of the customer j's preference for either vehicle p or vehicle q.

$$P^{(n)} = P_{(j,p,q)} = P\left(U_{jp} + \epsilon_{jp} > U_{jq} + \epsilon_{jq}\right) = \frac{e^{U_{jp}}}{e^{U_{jp}} + e^{U_{jq}}} \tag{6.2}$$

Parameter Estimation

We estimate the parameters of the logit model in Eq. (6.2) using conventional convex loss function minimization using the log-loss regularized with the l^2 norm.

$$\min_{\omega,\alpha} \frac{1}{N} \sum_{n=1}^{N} (y^{(n)} \log P^{(n)} + (1 - y^{(n)}) \log(1 - P^{(n)})) + \alpha \left\|\omega\right\|^2 \tag{6.3}$$

where $y^{(n)} = y_{(jpq)}$ is 1 if customer j chose vehicle p to purchase, and 0 if vehicle q was purchased; and α is the l^2 regularization hyperparameter. The optimization algorithm used to minimize this regularized loss function was stochastic gradient descent, with details of hyperparameter settings given in Section 6.5.

6.4 Feature Learning Models for Preference Prediction

We present three qualitatively different feature learning methods as introduced in Section 6.2: (1) principal component analysis, (2) low-rank + sparse matrix decomposition, and (3) exponential family sparse restricted Boltzmann machine. Furthermore, we discuss their extensions to better suit the market data described in Section 6.3, as well as derivation of theoretical guarantees.

6.4.1 Principal Component Analysis

Principal component analysis (PCA) maps the original data representation $\mathbf{x} = [x_1, x_2, \ldots, x_M]^T \in \mathbb{R}^{M\times 1}$ to a new feature representation $\mathbf{h} = [h_1, h_2, \ldots, h_K]^T \in \mathbb{R}^{K\times 1}$, $K \leq M$, with an orthogonal transformation $\mathbf{W} \in \mathbb{R}^{M\times K}$. Assume that the original data representation $\mathbf{x}$ has zero empirical mean (otherwise we simply subtract the empirical mean from $\mathbf{x}$). The mapping is given by:

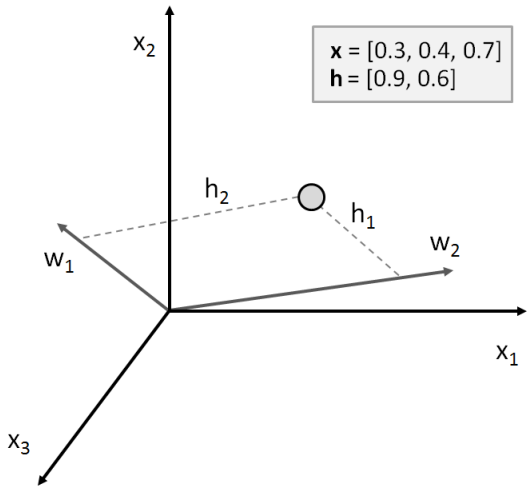


Figure 6.2: The concept of principal component analysis shown using an example with a data point represented by three original variables $\mathbf{x}$ projected to a two dimensional subspace spanned by $\mathbf{w}$ to obtain features $\mathbf{h}$.

$$\mathbf{h} = \mathbf{x}^T\mathbf{W} \tag{6.4}$$

The PCA representation has the following properties: (1) h_1 has the largest variance, and the variance of h_i is not smaller than the variance of h_j for all $j < i$; (2) the columns of $\mathbf{W}$ are orthogonal unit vectors; and (3) $\mathbf{h}$ and $\mathbf{W}$ minimize the reconstruction error ϵ:

$$\epsilon = ||\mathbf{x} - \mathbf{h}||^2 \tag{6.5}$$

When the q columns of $\mathbf{W}$ consist of the first q eigenvectors of $\mathbf{x}^T\mathbf{x}$, the above properties are all satisfied, and the PCA feature representation can be calculated by Equation (6.4). Since PCA is a projection onto a subspace, the features $\mathbf{h}$ in this case are not "higher order" functions of the original variables, but rather a linear mapping from original variables to a strictly smaller number of linear coefficients over the eigenvectors.

6.4.2 Low-Rank + Sparse Matrix Decomposition

The utility model U_{rp} given in Equation (6.1) can be rewritten into matrix form, in which Ω is a matrix reshaped from the "part-worth" coefficients vector ω:

$$U_{rp} = [\left(\mathbf{x}_c^{(j)}\right)^T, 1]\Omega\mathbf{x}_d^p \tag{6.6}$$

The decomposition of the original part-worth coefficients into a low-rank matrix and a sparse matrix may better represent customer purchase decisions than the large coefficient matrix of all pairwise interactions given in Equation ((6.1)) and as detailed in Section 6.2. Accordingly, we decompose Ω into a low-rank matrix $\mathbf{L}$ of rank r superimposed with a sparse matrix $\mathbf{S}$, i.e. $\Omega = \mathbf{L} + \mathbf{S}$. This problem may be solved in the general case exactly with the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{L},\mathbf{S}} \; l(\mathbf{L}, \mathbf{S}; \mathbf{X}_c, \mathbf{X}_d, \mathbf{y}) \\ s.t. \quad &rank(\mathbf{L}) \leq r \\ &\mathbf{S} \in \mathcal{C} \end{aligned} \tag{6.7}$$

where $\mathbf{X}_u$ and $\mathbf{X}_c$ are the full set of customer and vehicle data, $\mathbf{y}$ is the vector of whether customer j chose vehicle p or vehicle q, $l(\cdot)$ is the log-loss without the l^2 norm,

$$\begin{aligned} &l(\mathbf{L}, \mathbf{S}; \mathbf{X}_c, \mathbf{X}_d, \mathbf{y}) \\ &= \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} \log P^{(n)} + (1 - y^{(n)}) \log(1 - P^{(n)})) \end{aligned} \tag{6.8}$$

and $\mathcal{C}$ is a convex set corresponding to the sparse matrix $\mathbf{S}$. As this problem is intractable (NP-hard), we instead learn this decomposition of matrices using an ap-

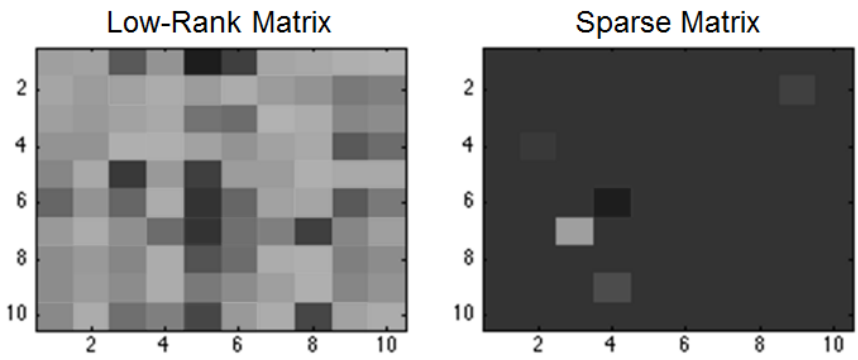


Figure 6.3: The concept of low-rank + sparse matrix decomposition using an example "part-worth coefficients" matrix of size 10 x 10 decomposed into two 10 x 10 matrices with low rank or sparse structure. Lighter colors represent larger values of elements in each decomposed matrix.

proximation obtained via regularized loss function minimization:

$$\min_{\mathbf{L},\mathbf{S}} \; l(\mathbf{L},\mathbf{S};\mathbf{X}_c,\mathbf{X}_d,\mathbf{y}) + \lambda_1||\mathbf{L}||_* + \lambda_2||\mathbf{S}||_1 \tag{6.9}$$

where $||\cdot||_*$ is the nuclear norm to promote low-rank structure, and $||\cdot||_1$ is the l_1-norm.

In particular, while a number of low-rank regularizations may be used to solve Eq. (6.9), e.g., trace norm and log-determinant norm (**?**). We choose the nuclear norm as it may be applied to any general matrix, while the trace norm and log-determinant regularization are limited to positive semidefinite matrices. Moreover, the nuclear norm is often considered optimal as $||\mathbf{L}||_*$ is the convex envelop of $Rank(\mathbf{L})$, implying that $||\mathbf{L}||_*$ is the largest convex function smaller than $Rank(\mathbf{L})$ (**?**).

Definition 1. For matrix $\mathbf{L}$,the nuclear norm is defined as,

$$||\mathbf{L}||_* := \sum_{i=1}^{\min(dim(L))} s_i(\mathbf{L})$$

where $s_i(\mathbf{L})$ is a singular value of $\mathbf{L}$.

6.4.2.1 Parameter Estimation

The non-differentiability of the convex low-rank + sparse approximation given in Eq. (6.9) necessitates optimizations techniques such as augmented Lagrangian (**?**), semi-definite programming (**?**), and proximal methods (**?**). Due to theoretical guarantees on convergence, we choose to train our model using proximal methods which are defined as follows.

Definition 2. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed proper convex function. The *proximal operator* of f is defined as

$$\mathbf{prox}_f(\mathbf{v}) = \arg\min_{\mathbf{x}} \left(f(\mathbf{x}) + \frac{1}{2}||\mathbf{v} - \mathbf{x}||_2^2 \right)$$

With these preliminaries, we now detail the proximal gradient algorithm used to solve Eq. (6.9) using low-rank and l^1 proximal operators. Denote $f(\cdot) = \lambda_1 || \cdot ||_*$, and its proximal operator as $prox_f$. Similarly denote the proximal operator for the l^1 regularization term by $prox_S$, $i = 1, \ldots n$.

With this notation, the proximal optimization algorithm to solve Equation ((6.9)) is given by Algorithm 1. Moreover, this algorithm is guaranteed to converge with constant step size as given by the following lemma (**?**).

Lemma 3. ***Convergence Property***

When ∇l is Lipschitz continuous with constant ρ, this method can be shown to converge with rate $O(\frac{1}{k})$ when a fixed step size $\eta_t = \eta \in (0, 1/\rho]$ is used. If ρ is not known, the step sizes η_t can be found by a line search; that is, their values are chosen in each step.

6.4.3 Restricted Boltzmann machine

The restricted Boltzmann machine (RBM) is an energy-based model in which an energy state is defined by a layer of M visible nodes corresponding to the original

Input: Data $\mathbf{X}_c$, $\mathbf{X}_d$, $\mathbf{y}$
Initialize $\mathbf{L}^0 = \mathbf{0}, \mathbf{S}^0 = \mathbf{0}$
repeat
  $\mathbf{L}^{t+1} = prox_f(\mathbf{L}^t - \eta_t \nabla_{\mathbf{L}^t} l(\mathbf{L}, \mathbf{S}; \mathbf{X}_c, \mathbf{X}_d, \mathbf{y}))$
  $\mathbf{S}^{t+1} = prox_S(\mathbf{S}^t - \eta_t \nabla_{\mathbf{S}^t} l(\mathbf{L}, \mathbf{S}; \mathbf{X}_c, \mathbf{X}_d, \mathbf{y}))$
until $\mathbf{L}^t$, $\mathbf{S}_i^t$ are converged

Algorithm 1: Low-Rank + Sparse Matrix Decomposition

variables $\mathbf{x}$ and a layer of K features denoted as $\mathbf{h}$. The energy for a given pair of original variables and features determines the probability associated with finding the system in that state; like nature, systems tend to states that minimize their energy and thus maximize their probability. Accordingly, maximizing the likelihood of the observed data $\mathbf{x}^{(1)} \dots \mathbf{x}^{(N)} \in \mathbb{R}^M$ and its corresponding feature representation $\mathbf{h}^{(1)} \dots \mathbf{h}^{(N)} \in \mathbb{R}^K$ is a matter of finding the set of parameters that minimize the energy for all observed data.

While traditionally this likelihood consists of binary variables and binary features, as described in Table 6.1 and Table 6.2, our passenger vehicle purchase data set consists of M_G Gaussian variables, M_B binary variables, and M_C categorical variables. We accordingly define three corresponding energy functions E_G, E_B, and E_C, in which each energy function connects the original variables and features via a weight matrix $\mathbf{W}$, as well as biases for each original variable and feature, $\mathbf{a}$ and $\mathbf{b}$ respectively.

Real-valued random variables (e.g., vehicle curb weight) are modeled using the Gaussian density. The energy function for Gaussian inputs and binary hidden nodes is:

$$\begin{aligned} E_G(\mathbf{x}, \mathbf{h}; \theta) = &- \sum_{m=1}^{M_G} \sum_{k=1}^{K} h_k w_{km} x_m \\ &+ \frac{1}{2} \sum_{m=1}^{M_G} (x_m - b_m)^2 - \sum_{k=1}^{K} a_k h_k \end{aligned} \tag{6.10}$$

where the variance term is clamped to unity under the assumption that the input data are standardized.

Binary random variables (e.g., gender) are modeled using the Bernoulli density.

The energy function for Bernoulli nodes in both the input layer and hidden layer is:

$$\begin{aligned} E_B(\mathbf{x}, \mathbf{h}; \theta) = &- \sum_{m=1}^{M_B} \sum_{k=1}^{K} h_k w_{km} x_m \\ &- \sum_{m=1}^{M_B} x_m b_m - \sum_{k=1}^{K} a_k h_k \end{aligned} \tag{6.11}$$

Categorical random variables (e.g., vehicle manufacturer) are modeled using the categorical density. The energy function for categorical inputs with Z_m classes for m-th categorical input variable (e.g., Toyota, General Motors, etc.) is given by:

$$\begin{aligned} E_C(\mathbf{x}, \mathbf{h}; \theta) = &- \sum_{m=1}^{K_m} \sum_{k=1}^{K} \sum_{z=1}^{Z_m} h_k w_{kmz} \delta_{mz} x_{mz} \\ &- \sum_{m=1}^{M_C} \sum_{z=1}^{Z_m} \delta_{mz} x_{mz} b_{mz} - \sum_{k=1}^{K} a_k h_k \end{aligned} \tag{6.12}$$

where $\delta_{mz} = 1$ if $x_{mz} = 1$ and 0 otherwise.

Given these energy functions for the heterogeneous original variables, the probability of a state with energy $E(\mathbf{x}, \mathbf{h}; \theta) = E_G(\mathbf{x}, \mathbf{h}; \theta) + E_B(\mathbf{x}, \mathbf{h}; \theta) + E_C(\mathbf{x}, \mathbf{h}; \theta)$, in which $\theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$ are the energy function weights and bias parameters, is defined by the Boltzmann distribution.

$$P(\mathbf{x}, \mathbf{h}) = \frac{e^{-E(\mathbf{x}, \mathbf{h}; \theta)}}{\sum_{\mathbf{x}} \sum_{\mathbf{h}} e^{-E(\mathbf{x}, \mathbf{h}; \theta)}} \tag{6.13}$$

The "restriction" on the RBM is to disallow visible-visible and hidden-hidden node connections. This restriction results in conditional independence of each individual hidden unit h given the vector of inputs $\mathbf{x}$, and each visible unit x given the vector of hidden units $\mathbf{h}$.

$$P(\mathbf{h}|\mathbf{x}) = \prod_{n=1}^{N} P(h_n|\mathbf{x}) \tag{6.14}$$

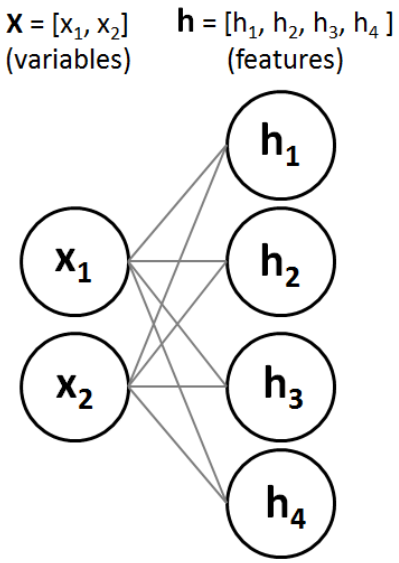


Figure 6.4: The concept of the exponential family sparse restricted Boltzmann machine. The original data are represented by nodes in the visible layer by $[x_1, x_2]$, while the feature representation of the same data is represented by nodes in the hidden layer $[h_1, h_2, h_3, h_4]$. Undirected edges are restricted to being only between the original layer and the hidden layer, thus enforcing conditional independence between nodes in the same layer.

$$P\left(\mathbf{x}|\mathbf{h}\right) = \prod_{k=1}^{K} P\left(x_k|\mathbf{h}\right) \tag{6.15}$$

The conditional density for a single binary hidden unit given the combined K_G Gaussian, K_B binary, and K_C categorical input variables is then:

$$\sigma(a_n + \sum_{k=1}^{K_G} w_{nk}x_k + \sum_{k=1}^{K_B} w_{nk}x_k + \sum_{k=1}^{K_C}\sum_{d=1}^{D_k} w_{nk}\delta_{kd}x_k) \tag{6.16}$$

where $\sigma(s) = \frac{1}{1+\exp(-s)}$ is a sigmoid function.

For an input data point $\mathbf{x}^{(n)}$, its corresponding feature representation $\mathbf{h}^{(n)}$ is given by sampling the "activations" of the hidden nodes.

$$\left[P(h_1 = 1|\mathbf{x}, \theta)\,,\, \ldots\,,\, P(h_N = 1|\mathbf{x}, \theta)\right] \tag{6.17}$$

Parameter Estimation

To train the model, we optimize the weight and bias parameters $\theta = \{\mathbf{W}, \mathbf{b}, \mathbf{a}\}$ by minimizing the negative log-likelihood of the data $\{\mathbf{x}^{(1)} \dots \mathbf{x}^{(N)}\}$ using gradient descent. The gradient of the log-likelihood is:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\sum_{n=1}^{N}\log P\left(\mathbf{x}^{(n)}\right) &= \frac{\partial}{\partial\theta}\sum_{n=1}^{N}\log\sum_{\mathbf{h}}P\left(\mathbf{x}^{(n)},\mathbf{h}\right) \\
&= \frac{\partial}{\partial\theta}\sum_{n=1}^{N}\log\sum_{\mathbf{h}}\frac{e^{-E(\mathbf{x}^{(n)},\mathbf{h})}}{\sum_{\mathbf{x},\mathbf{h}}e^{-E(\mathbf{x}^{(n)},\mathbf{h})}} \\
&= \sum_{n=1}^{N}\mathbb{E}_{\mathbf{h}|\mathbf{x}^{(n)}}\left[\frac{\partial}{\partial\theta}E\left(\mathbf{x}^{(n)},\mathbf{h}\right)\right] \\
&\quad - \mathbb{E}_{\mathbf{h},\mathbf{x}}\left[\frac{\partial}{\partial\theta}E\left(\mathbf{x},\mathbf{h}\right)\right]
\end{aligned} \tag{6.18}$$

The gradient is the difference of two expectations, the first of which is easy to compute since it is "clamped" at the input datum $\mathbf{x}$, but the second of which requires the joint density over the entire $\mathbf{x}$ space for the model.

In practice, this second expectation is approximated using the contrastive divergence algorithm by Gibbs, sampling the hidden nodes given the visible nodes, then the visible nodes given the hidden nodes, and iterating a sufficient number of steps for the approximation (**?**). During training, we induce sparsity of the hidden layer by setting a target activation β_k, fixed to 0.1, for each hidden unit h_k (**?**). The overall objective to be minimized is then the negative log-likelihood from Equation (6.18) and a penalty on the deviation of the hidden layer from the target activation. Since the hidden layer is made up of sigmoid densities, the overall objective function is:

$$\begin{aligned}
&\sum_{n=1}^{N}\log\sum_{\mathbf{h}}P\left(\mathbf{x}^{(n)},\mathbf{h}\right) \\
&+\lambda_3\sum_{k=1}^{K}\left(\beta_k^{(n)}\log h_k + \left(1-\beta_k^{(n)}\right)\log\left(1-h_k\right)\right),
\end{aligned} \tag{6.19}$$

where λ_3 is the hyperparameter trading off the sparsity penalty with the log-likelihood.

6.5 Experiment

The goal in this experiment was to assess how preference prediction accuracy changes when using the same preference model on three different representations of the same data set. The preference model used, as discussed in Section 6.3, was the l^2 logit, while the three representations were the original variables, low-rank + sparse features, and RBM features. The same experimental procedure was run on each of these three representations, where the first representation acts as a baseline for prediction accuracy, and the next two representations demonstrate the relative gain in preference prediction accuracy when using features.

In addition, we performed an analysis of how the hyperparameters affected design preference prediction accuracy for the hyperparameters used in the PCA, LSD, and RBM feature learning methods. For PCA, the hyperparameter was the dimensionality K of the subspace spanned by the eigenvectors of the PCA method. For LSD, the hyperparameters were the rank penalty λ_1, which affects the rank of the low-rank matrix L, and the sparsity penalty λ_2, which influences the number of non-zero elements in the sparse matrix S, both found in Equation (6.9). For RBM, the hyperparameters were the sparsity penalty λ_3, which controls the number of features activated for a given input datum, and the overcompleteness factor γ, which defines by what factor the dimensionality of the feature space is larger than the dimensionality of the original variable space, both of which are found in Equation (6.19).

The detailed experiment flow is summarized below and illustrated in Figure 6.5:

1. The raw choice data set of pairs of customers and purchased designs, described in Section 6.3, was randomly split 10 times into 70% training, 10% validation, and 20% test sets. This was done in the beginning to ensure no customers in

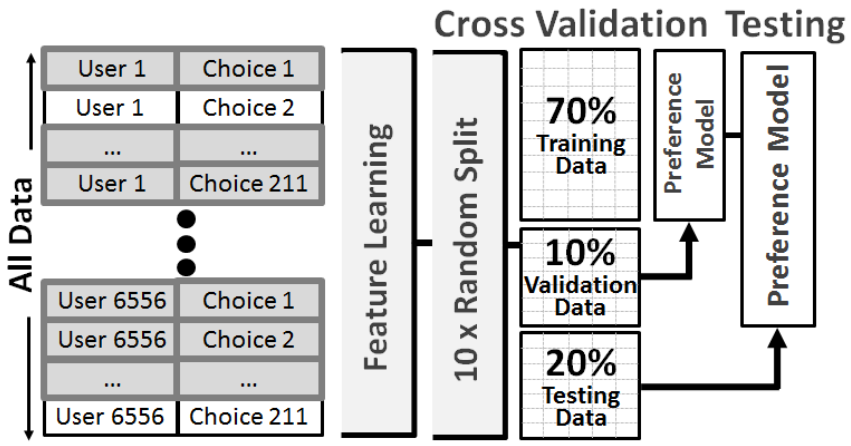


Figure 6.5: Data processing, training, validation, and testing flow.

the training sets ever existed in the validation or test sets.

2. Choice sets were generated for each training, validation, and test sets for all 10 randomly shuffled splits as described in Section 6.3. This process created a training data set of 832,000 data points, a validation data set of 104,000 data points, and a testing data set of 225,056 data points, for each of the 10 shuffled splits.

3. Feature learning was conducted on the training sets of customer variables and vehicle variables for a vector of 5 different values of K for PCA features, a grid of 25 different pairs of low-rank penalty λ_1 and sparsity penalty λ_2 for the LSD features, and a grid of 56 different pairs of sparsity λ_3 and overcompleteness γ hyperparameters for RBM features. For PCA features, these hyperparameters were $K \in \{30, 50, 70, 100, 150\}$. For LSD features, these hyperparameters were $\lambda_1 \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$ and $\lambda_2 \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$. For RBM, these hyperparameters were $\lambda_3 \in \{4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0\}$ and $\gamma \in \{0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0\}$. These hyperparameter settings were selected by pilot testing large ranges of parameter settings to find relevant regions for upper and lower hyperparameter bounds, with numbers of hyperpa-

Design Preference Model	Feature Representation	Prediction Accuracy (std. dev.) (ρ-value) $N = 10,000$	Prediction Accuracy (std. dev.) (ρ-value) $N = 1,161,056$
Logit Model	Original Variables (No Features)	69.98% (1.82%) (N/A)	75.29% (0.98%) (N/A)
Logit Model	Principle Component Analysis	61.69% (1.24%) (1.081e-7)	62.03% (0.89%) (8.22e-10)
Logit Model	Low-Rank + Sparse Matrix Decomposition	**76.59% (0.89%) (3.276e-8)**	**77.58% (0.81%) (4.286e-8)**
Logit Model	Exponential Family Sparse RBM	74.99% (0.64%) (2.3e-5)	75.15% (0.81%) (0.136)

Table 6.3: Averaged preference prediction accuracy on held-out test data using the logit model with the original variables or the three feature representations. Average and standard deviation were calculated from 10 random training and testing splits common to each method, while test parameters for each method were selected via cross validation on the training set.

rameters selected based on computational constraints.

4. Each of the validation and testing data sets were encoded using the feature learning methods learned for each of the 5 PCA hyperparameters K, 25 (λ_1, λ_2) LSD hyperparameter pairs, and 56 (λ_3, γ) RBM hyperparameter pairs.

5. The encoded feature data was combined with the original variable data in order to separate linear term effects of the original variables with higher order effects from the features. While this introduces a degree of information redundancy between features and original variables, the regularization term in Equation 6.3 mitigates effects of collinearity. Each datum consists of the features concatenated with the original variables, then input into the bilinear utility model. Specifically, for some customer features $\mathbf{h}_u$ and customer variables $\mathbf{x}_u$, we used $\mathbf{h}_{u'}^T := [\mathbf{x}_u^T, \mathbf{h}_u^T]$ to define the new representation of the customer; likewise, for

some vehicle features $\mathbf{h}_c$ and vehicle variables $\mathbf{x}_c$, we used $\mathbf{h}_{c'}^T := [\mathbf{x}_c^T, \mathbf{h}_c^T]$ to define the new representation of the customer. Combined with Equation (6.1), a single data point used for training is the difference in utilities between vehicle p and vehicle q for a given customer r.

$$\left[\mathbf{h}_{u'}^{(r)} \otimes \left(\mathbf{h}_{c'}^{(p)} - \mathbf{h}_{c'}^{(q)}\right), \mathbf{h}_{c'}^{(p)} - \mathbf{h}_{c'}^{(q)}\right] \tag{6.20}$$

Note that the dimensionality of each datum could range above 100,000 dimensions for the largest values of γ.

6. For each of these training sets, 6 logit models were trained in parallel over minibatches of the training data, corresponding to 6 different settings of the l^2 regularization parameter $\alpha = 0.00001, 0.0001, 0.001, 0.01, 0.1, 1.0$. These logit models were optimized using stochastic gradient descent, with learning rates inversely related to the number of training examples seen (**?**).

7. Each logit model was then scored according to its respective held-out validation data set. The hyperparameter settings ($\alpha_{BASELINE}$) for the original variables, (K_{PCA}, α_{PCA}) for PCA feature learning, (λ_1, λ_2) for LSD feature learning, and ($\lambda_3, \gamma, \alpha_{RBM}$) for RBM feature learning with the best validation accuracy were saved. For each of these four sets of best hyperparameters, Step 3 was repeated to obtain the set of corresponding features on each of the 10 random shuffled training plus validation sets.

8. Logit models corresponding to the baseline, PCA features, LSD features, and RBM features were retrained for each of the 10 randomly shuffled and combined training and validation. The prediction accuracy for each of these 10 logit models was assessed on the corresponding "held out" test sets in order to give average and standard deviations of the design preference predictive accuracy

for the baseline, PCA features, LSD features, and RBM features.

6.6 Results

Table 6.3 shows the averaged test set prediction accuracy of the logit model using the original variables, PCA features, LSD features, and RBM features. Prediction accuracy averaged over 10 random training and held-out testing data splits are given, both for the partial data $N = 10,000$ and the full data $N = 1,161,056$ cases. Furthermore, we include the standard deviation of the prediction accuracies and a 2-sided t-test relative to the baseline accuracy for each feature representation.

The logit model trained with LSD features achieved the highest predictive accuracy on both the partial and full data sets, at 76.59% and 77.58%, respectively. This gives evidence that using features can improve design preference prediction accuracy as the logit model using the original variables achieved an averaged accuracy of 69.98% and 75.29%, respectively. The improvement in design preference prediction accuracy is greatest for the partial data case, as evidenced by both the LSD and RBM, yet the improvement with the full data case shows that the LSD feature learning method is still able to improve prediction accuracy within the capacity of the logit model. The RBM results for the full data case do not show significant improvement in prediction accuracy. Finally, we note a relative loss in design preference prediction accuracy when using PCA as a feature learning method, both for the partial and full data sets, suggesting the heavy assumptions built into PCA are overly restrictive.

The parameter settings for the LSD feature learning method give additional insight to the preference prediction task. In particular, the optimal settings of λ_1 and λ_2 obtained through cross validation on the 10 random training sets was ranged from $r = 29$ to $r = 31$. This significantly reduced rank of the part-worth coefficient matrix given in Eq. (6.1) suggests that the vast majority of interactions between customer variables and design variables given in Table 6.1 and Table 6.2 do not significantly

contribute to overall design preferences. This insight allows us to introspect into important feature pairings on a per-customer basis to inform design decisions.

We have shown that even "simple" single-layer feature learning can significantly increase predictive accuracy for design preference modeling. This finding signifies that features more effectively capture the design preferences than the original variables, as features form functions of the original variables more representative of the customer's underlying preference task. This offers designers opportunity for new insights if these features can be successfully interpreted and translated to actionable design decisions; however, given the relatively recent advances in feature learning methods, interpretation and visualization of features remains an open challenge–see Section 6.7 for further discussion.

Further increases to prediction accuracy might be achieved by stacking multiple feature learning layers, often referred to as "deep learning". Such techniques have recently shown impressive results by breaking previous records in image recognition by large margins (**?**). Another possible direction for increasing prediction accuracy may be in developing novel architectures that explicitly capture the conditional statistical structure between customers and designs. These efforts may be further aided through better understanding of the limitations of using feature learning methods for design and marketing research. For example, the large number of parameters associated with feature learning methods results in greater computational cost when performing model selection; in addition to the cross-validation techniques used in this paper, model selection metrics such as BIC and AIC may give further insight along these lines.

6.7 Using Features for Design

Using features can support the design process in at least two directions: (1) Features interpretation can offer deeper insights into customer preferences than the orig-

inal variables, and (2) feature visualization can lead to a market segmentation with better clustering than with the original variables. These two directions are still open challenges given the relative nascence of feature learning methods. Further investigation is necessary to realize the above design opportunities and to justify the computational cost and implementation challenges associated with feature learning methods.

The interpretation and visualization methods may be used with conventional linear discrete choice modeling (e.g., logit models). However, deeper insights are possible through interpreting and visualizing features, assuming that features capture more effectively the underlying design preference prediction task of the customer as shown through improved prediction accuracy on held-out data. Since we are capturing "functions" of the original data, we are more likely to interpret and visualize feature pairings such as "eco-friendly" vehicle and "environmentally conscious" customer; such pairing may ultimately lead to actionable design decisions.

6.7.1 Feature Interpretation of Design Preferences

Similar to PCA, LSD provides an approach to interpret the learned features by looking at the linear combinations of original variables. The major difference between features learned using PCA versus LSD is their different linear combinations; in particular, features learned by LSD are more representative as they contain information from both the data distribution and the preference task, while PCA features only contain information from the data distribution.

As introduced in section 6.4.2, the weight matrix Ω is decomposed into a low-rank matrix $\mathbf{L}$ and a sparse matrix $\mathbf{S}$, i.e. $\Omega = \mathbf{L} + \mathbf{S}$. The nonzero elements in the sparse matrix $\mathbf{S}$ may be interpreted as the weight of the product of its corresponding original design variables and customer variables. As for the low-rank matrix $\mathbf{L}$, features can be extracted by linearly combining the original variable according to the singular value decomposition (SVD) for $\mathbf{L}$. The singular value decomposition is a factorization of

the $(m+1) \times n$ matrix $\mathbf{L}$ in the form $\mathbf{L} = \mathbf{U}\Sigma\mathbf{V}$, where $\mathbf{U}$ is a $(m+1) \times (m+1)$ unitary matrix, Σ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers $\sigma_1, \sigma_2, \ldots, \sigma_{\min(m+1,n)}$ on the diagonal, and $\mathbf{V}$ is a $(n) \times (n)$ unitary matrix. Rewriting Equation (6.6):

$$\begin{aligned} U_{rp} &= \left[\left(\mathbf{x}_c^{(j)}\right)^T, 1\right] \mathbf{L}\mathbf{x}_d^p + \left[\left(\mathbf{x}_c^{(j)}\right)^T, 1\right] \mathbf{S}\mathbf{x}_d^p \\ &= \left[\left(\mathbf{x}_c^{(j)}\right)^T, 1\right] \mathbf{U}\Sigma\mathbf{V}\mathbf{x}_d^p + \left[\left(\mathbf{x}_c^{(j)}\right)^T, 1\right] \mathbf{S}\mathbf{x}_d^p \\ &= \sum_{i=1}^{min(m+1,n)} \sigma_i \left[\left(\mathbf{x}_c^{(j)}\right)^T, 1\right] \mathbf{u}_i\mathbf{v}_i\mathbf{x}_d^p + \left[\left(\mathbf{x}_c^{(j)}\right)^T, 1\right] \mathbf{S}\mathbf{x}_d^p \end{aligned} \tag{6.21}$$

where $\mathbf{u}_i$ is the i-th column of matrix U, and $\mathbf{v}_i$ is the i-th row of matrix V. The i-th user feature $\left[\left(\mathbf{x}_c^{(j)}\right)^T, 1\right] \mathbf{u}_i$ is a linear combination of original user variables; the i-th design feature $\mathbf{v}_i\mathbf{x}_d^p$ is a linear combination of original design variables; and σ_i represents the importance of this pair of features for the customer's design preferences.

Interpreting these features in the vehicle preference case study, we found that the most influential feature pairing (i.e., largest σ_i) corresponds to preference trends at the population level: Low price but luxury vehicles are preferred, and Japanese vehicles receive the highest preference while GM vehicles receive the lowest preference. The second most influential feature pairing represents a rich customer group, with preferred vehicle groups being both expensive and luxurious. The third most influential feature pairing represents an elder user group, with their preferred vehicles as large but with low net horsepower.

6.7.2 Features Visualization of Design Preferences

We now visualize features to understand what insights for design decision making. Specifically, we make early-stage inroad to visual market segmentation performed in an estimated feature space, thus clustering customers in a representation that better captures their underlying design preference decisions.

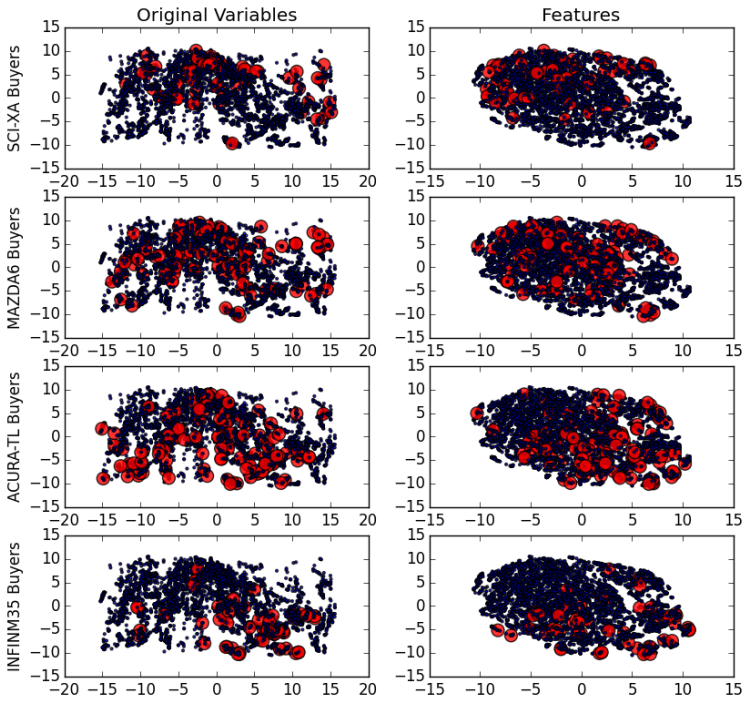


Figure 6.6: Optimal vehicle distribution visualization. Every point represents the optimal vehicle for one consumer. In the left column, the optimal vehicle is inferred using the utility model with original variables. In the right column, LSD features are used to infer the optimal vehicle. In the first row, the optimal vehicles from SCI-XA customers are marked in big red points. Similarly, the optimal vehicles from MAZDA6, ACURA-TL and INFINM35 customers are marked in big red points respectively.

We begin by looking at the utility model U_{rp} given in Equation (6.1) and note that the inner product between Ω and the variables $\mathbf{x}_u^{(r)}$ representing customer r may be interpreted as customer r's optimal vehicle, denoted $\mathbf{x}_{opt}^{(r)}$:

$$\mathbf{x}_{opt}^{(r)} = \left(\mathbf{x}_u^{(r)}\right)^T \Omega_{out} + \mathbf{1}^T \Omega_{main} \tag{6.22}$$

where Ω_{out} is the matrix reshaped from the coefficients of Ω corresponding to the outer product given in Equation (6.1), Ω_{main} is the matrix reshaped from the remaining coefficients, and $\mathbf{1}$ is a vector consisting of 1's with the same dimension as $\mathbf{x}_u^{(r)}$. We rewrite the utility model U_{rp} given in Equation (6.1) in terms of the optimal vehicle $\mathbf{x}_{opt}^{(r)}$:

$$U_{rp} = \left(\mathbf{x}_{opt}^{(r)}\right)^T \mathbf{x}_d^p \tag{6.23}$$

According to the geometric meaning of inner product, the smaller the angle between $\mathbf{x}_d^p$ and $\mathbf{x}_{opt}^{(r)}$ is, the larger will be the utility U_{rp}. In this way, we have an interpretable method of improving upon the actual purchased vehicle design in the form of an 'optimal' vehicle vector. This optimal vehicle vector could be useful for a manufacturer developing a next-generation design from a current design, particularly as the manufacturer would target a specific market segment.

We now provide a visual demonstration of using an optimal vehicle derived from feature learning to suggest a design improvement direction. First, we calculate the optimal vehicle using Equation (6.22) for every customer in the data set. Then, we visualize these optimal vehicle points by reducing their dimension using t-distributed stochastic neighbor embedding (t-SNE), an advanced nonlinear dimension reduction technique that embeds similar objects into nearby points (van der Maaten, 2008). Finally, optimal vehicles from targeted market segments are marked in red.

Figure 6.6 shows the optimal vehicles for the SCI-XA, MAZDA6, ACURA-TL and INFINM35 customer groups using red points respectively. We observe that the optimal vehicle moves from the left-top corner to the right-bottom corner as the purchased vehicles become more luxurious using the LSD features, while the optimal vehicles in the original variable representation show overlap, especially for MAZDA6 and ACURA-TL customers. In other words, we are visualizing what has been shown quantitatively through increased preference prediction accuracy; namely, that optimal vehicles represented using LSD features as opposed to the original variables result in a larger separation of various market segments' optimal vehicles.

The contribution of this demonstration is not the particular introspection on the chosen example with MAZDA6 and ACURA-TL customers. Instead, this demonstration is significant as it suggests it is possible to perform feature-based market

segmentation purely using visual analysis. Such visual analysis is likely to be more useful to practicing designers and marketers, as it abstracts away the underlying mathematical mechanics of feature learning.

6.8 Summary

Feature learning is a promising method to improve design preference prediction accuracy without changing the design preference model or the data set. This improvement is obtained by transforming the original variables to a feature space acting as an intermediate step as shown in Figure 6.1. Thus, feature learning complements advances in both data gathering and design preference modeling.

We presented three feature learning methods–principal component analysis, low-rank plus sparse matrix decomposition, and sparse exponential family restricted Boltzmann machines–and applied them to a design preference data set consisting of customer and passenger vehicle variables with heterogeneous unit types, e.g., gender, age, # cylinders.

We then conducted an experiment to measure design preference prediction accuracy involving 1,161,056 data points generated from a real purchase dataset of 5582 customers. The experiment showed that feature learning methods improve preference prediction accuracy by 2-7% for a small and full dataset, respectively. This finding is significant, as it shows that features offer a better representation of the customer's underlying design preferences than the original variables. Moreover, the finding shows that feature learning methods may be successfully applied to design and marketing data sets made up of variables with heterogeneous data types; this is a new result as feature learning methods have primarily been applied on homogeneous data sets made up of variables of the same distribution.

Feature interpretation and visualization offer a promise for using features to support the design process. Specifically, interpreting features can give designers deeper

insights of the more influential pairings of vehicle features and customer features, while visualization of the feature space can offer deeper insights when performing market segmentation. These new findings suggest opportunities to develop feature learning algorithms that are not only more representative of the customer preference task as measured by prediction accuracy but also easier to interpret and visualize by a domain expert. Methods allowing easier interpretation of features would be valuable when translating the results of more sophisticated feature learning and preference prediction models into actionable design decisions.

CHAPTER VII

Conclusion

7.1 Summary of Dissertation

Crowdsourcing for engineering design is an approach for obtaining human input across a number of evaluators or customers separate from the designer herself, enabled by the reach and scale of the internet and modern computational processing, for a given objective or subjective design decision. This approach has seen much recent attention at industrial companies and governmental agencies, as it offers the opportunity to make good design decisions and to catch bad design decisions at the early stages of the design process, thus saving on cost and time overruns that often plague complex engineering design.

While a number of successes has been qualitatively documented via both business case studies and academic literature from the product innovation and management communities, these successes are primarily related to simple tasks, e.g., image annotation. In contrast, crowdsourcing for complex tasks associated with engineering design decisions have often been unsuccessful due to the heterogeneity of evaluator expertise or customer preference within the crowd. These qualitative findings run parallel to the observation that this lack of quantitative models to appropriately model this heterogeneity for engineering design.

Chapter 1 discussed these aforementioned qualitative findings and observations,

leading to the research gap studied in the dissertation. We then introduced the framework for the following chapters, namely, the spectrum of expertise necessary for a given design task as was given in Figure 1.6.

Chapter 2 investigated the case in which only a small minority of the crowd has sufficient expertise for accurate evaluation, and showed that a Bayesian network model, qualitatively similar to other "off-the-shelf" crowd aggregation models, failed to combine the crowd's evaluations into an accurate combined evaluation due to the relatively few experts in the crowd. Most importantly, it was found that this failure was due to the relatively few experts being overshadowed by numerous clusters of "consistent, yet incorrect" evaluators.

Chapter 3 aimed to "identify the experts" in the crowd, yet was not able to use commonly prescribed variables such as demographics, reaction times, or a number of benchmark tests of mechanical reasoning as identifiers of expertise. Instead, we were only able to find the experts using a "simple version" of the actual hard version of the design task, thus allowing successful crowd aggregation by filtering out non-experts.

Chapter 4 moved to the case of a subjective design decision, the balance between brand recognition and design freedom of the aesthetic styling of concept vehicle designs, a design decision that incorporated expertise in the form of how well a customer could recognize previous models of a design. Heterogeneity of expertise in this case was used to filter non-experts, to successfully aggregate customer preferences across various brands.

Chapter 5 examined the case of very low expertise required for design decisions, specifically preferences over the visual fidelity of generated 2D images of vehicle designs in a design space estimated using deep generative models. In this case, the single crowd-level preference acted as an optimizer for parameters and model architecture decisions governing the resulting visual fidelity of the generated designs.

Chapter 6 then looked at the case of no expertise needed for the design task,

in which every individual was assumed to perfectly know his or her preferences as elicited by actual design purchase decisions. In this case, capturing the heterogeneity of preference resulted in improvements in design purchase prediction, thus enabling more successful crowd aggregation models for design.

7.2 Contribution to Design Science

The main research contribution of this work is a systematic quantitative study across the spectrum of heterogeneity of evaluator expertise or customer preferences within the crowd, leading to quantitative understanding of why crowdsourcing systems have often been unsuccessful for industrial companies and governmental agencies engaged in engineering design. This contribution may be expanded as follows:

(i) We have quantitatively investigated and characterized clusters of heterogeneity found within the crowd for both objective evaluations and subjective preferences. In the objective case, these clusters led to "consistent, yet incorrect" evaluators, which washed out the crowd consensus. In the subjective case, these clusters led to difficulty in suggesting optimal designs, as these clusters were more appropriately represented as functions of the original variables themselves in the form of known design attributes and unknown design features.

(ii) We have introduced probabilistic models of crowd aggregation that mitigate the issues of heterogeneity across this spectrum from objective evaluations to subjective preferences. In the objective case, this included incorporating discriminative information about the evaluator's expertise in the form of performance on a simple version of the difficult design task. In the subjective case, this included explicitly modeling heterogeneity of design attributes and design features.

(iii) We have made inroads to visual analysis tools that abstract away the underlying mathematics, thus allowing more practical usage by designers within

enterprises–be they industrial designers, marketers, or executive strategists.

This main contribution is supported by five chapters of research spanning a spectrum from objective design decision requiring high expertise (i.e., only a few experts exist in the crowd) to subjective design decisions requiring no expertise (i.e., everyone is an expert of their own preferences).

7.2.1 Limitations

A number of limitations exist in our research findings. First, we will discuss two major limitations common to all chapters in this dissertation, then discuss limitations specific to each chapter.

The first limitation common to all chapters is that we are only dealing with a very specific type of crowdsourcing—offline, static, non-collaborative, non-active, incentivized by fiscal payment, and other properties as detailed in Figure 1.2. There are a very large number of other types of crowdsourcing, with many ongoing research challenges and active research programs both within and outside of the design research community (e.g., see Panchal (2015c) or gamification Ren *et al.* (2015a)).

The second limitation common to all chapters is that we have assumed design representations that are often very restrictive in eliciting an evaluator or customer's true visceral and perceptual responses. These representations have often been in the form of 2D images or 3D meshes. Contrast these representations with those used at industrial companies, e.g., partial and full scale vehicle concept designs.

Chapter 2 had the following additional limitations: (1) Evaluators evaluate designs without systematic biases, i.e., given infinite chances of evaluating one specific design, the average score of the evaluators will converge to the true score of that design regardless of their expertise Nunnally & Bernstein (2010); Caragiannis *et al.* (2013); note that this assumption also implies that no evaluators purposely give bad evaluations; (2) evaluations are independent, i.e., the evaluation on one design from

one evaluator will not be affected by the evaluation made by that evaluator for any other design, nor will it be affected by the evaluation given by a different evaluator; (3) the expertise of evaluators is constant during the entire evaluation process; (4) all evaluators are fully incentivized and do not exhibit fatigue. Without loss of generality, we consider human evaluations real-valued in the range of zero to one.

Chapter 3 assumed that all evaluations are noiseless and occur at the same time for the same engineering design task. These assumptions are not strictly valid for real engineering design tasks in the workplace due to inter-team communication issues Austin-Breneman *et al.* (2014); McGowan *et al.* (2013), as noted in Section 1.2. We also did not account for evaluator learning during the evaluation process. Such task learning has been shown to significantly affect crowdsourced evaluation Wu & Duffy (2004).

The analysis involved linear models or, more accurately, assumptions stemming from linear models. In particular, we used correlation coefficients and hypothesis tests stemming from Gaussian assumptions on the data. While these assumptions are limiting, the very low linear correlation amongst assumed independent variables suggests that it is unlikely that we may be seeing all variation in the data contained in high-order moments or amongst statistical dependencies involving joint variation of multiple variables.

Chapter 4 has limitations in that the design space spanned by the parameterization of geometric variables for the 3D models does not capture the entire set of possible vehicle face design concepts. While this is in part why we assumed brand recognition as a linear function of attributes — and attributes as an implicit nonlinear function of geometric variables— future studies may greatly differ in their parameterizations.

Filtering the data for brand-conscious customers has also some limitations. We assumed that brand recognition accuracy is a static quantity throughout the survey. This does not account for familiarity with the brands after consistently seeing the

same four images throughout the survey. Further, a larger number of data points would reduce the uncertainty in Figure 4.4, as well as allow filtering on customers with higher average brand recognition accuracy over current MY2014 vehicles.

Further limitations to Chapter 4 include the following for the crowdsourced function estimation approach: First, attribute values will change depending on which cars are involved in the ranking. Second, the formulation assumes that customers are homogeneous in their perceptions of the design attributes. While this assumption is certainly not always true, we mitigate the effect of heterogeneity by normalizing for the relative contribution of a design attribute to either design freedom or brand recognition as given in Eq. (4.2). Finally, we note that including heterogeneity in customer responses to design attributes may significantly increase fidelity of the brand recognition prediction model. Such heterogeneity may be captured using models that incorporate clustering formulations or formulations that impose deviations from a common crowd prior distribution Evgeniou *et al.* (2007b); Abernethy *et al.* (2008).

Note also that Chapter 4 only considered designs from MY2014, limiting these static findings from time-series trends. Future work considering design data over a number of years would provide additional insight as brands and design languages often undergo dramatic shifts Ma *et al.* (2014); Tucker & Kim (2011). Furthermore, this study considered only luxury brands, in part because such brand imagery tends to be more recognizable. Insights into whether the same findings and methodology are appropriate for non-luxury brands would be interesting to explore. Further, design domains besides automotive offer additional opportunities for exploration.

Chapter 5 had a number of limitations as discussed textually and visually in Section 5.5.

Chapter 6 had limitations primarily on the method of estimating heterogeneity in the crowd. In particular, we assumed a bilinear model of utility. Although we use both linear feature learning (low-rank matrix decomposition) and nonlinear feature

learning (exponential family restricted Boltzmann machine) to transform the original variables before inputting them into the utility model, this bilinear form to capture heterogeneity is rather naïve. We discuss this further in future work below.

7.2.2 Future Work

There are three major directions of future work that may prove valuable to design science, the first and second directions being more theoretical while the third being more practical. The first major direction of future work is rigorously formalizing the findings of this dissertation via bounds on when crowdsourcing systems fail for a given design decision. While we have identified the reason why, and even intuition on the distribution of expertise or preferences for design tasks, we have not rigorously proved any of these conditions. Rigorous proofs, provided they are useful, may further advance the practicality of crowdsourcing systems for enterprises engaged in engineering design.

Much theoretical work has proved bounds on regret within a model class or convergence rates towards optimal models within an assumed class; however, these bounds are not clearly useful for practical crowdsourcing systems. In other words, since these bounds often assume infinite data (e.g., Chernoff bound), they may actually give misleading input to actionable design decisions. Recent work however in finite-sample bounds has shown success for the Dawid-Skene crowd aggregation models Li & Yu (2014), as introduced in Table 1.1. This area of research likely has the potential to improve practical crowdsourcing systems.

The second major direction of future work is in better characterization of the heterogeneous clusters of expertise or preferences. Specifically, cluster properties such as cluster shape distribution, size distribution, hard boundary vs. distributed membership, and whether clusters hierarchical vs. non-hierarchical are important to understand. These cluster properties are necessarily crowd and design decision depen-

dent, suggesting tools that infer these properties, perhaps in conjunction with expert designers, may prove valuable. Once such cluster properties are better characterized, they may be incorporated into crowd aggregation models that build on the models developed in this dissertation.

The third major direction of future work is in more advanced visual analysis tools for practicing designers at industrial companies and governmental agencies. After numerous meetings with practicing designers at both these enterprises, it is readily apparent that any crowdsourcing system must be user friendly by abstracting away the underlying appropriate yet unwieldy mathematics.

While we have made a preliminary inroad to to practical and easy-to-use visual tools more suited to designers—be they industrial designers, marketers, or executive strategists—we are only scratching the surface of what a production-level crowdsourcing system may require. We imagine any crowdsourcing system must give designers the opportunity to visually and interactively select expert clusters in the objective case, or market segments in the subjective case.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ordinal Representaitons and Uniqueness of Preference Orderings. 00000.

2014a. *Amazon Mechanical Turk*. 00082.

2014b. *EvoStock.com*. 00000.

2014c. *General Motors Brand Equity Research*. Internal. General Motors, Warren, MI. 00000.

Aaker, David A. 2009. *Managing Brand Equity*. New York, NY: Simon and Schuster. 08613.

Aaker, David A., & Keller, Kevin Lane. 1990. Consumer Evaluations of Brand Extensions. *The Journal of Marketing*, **54**(1), 27–41. 02916.

Abernethy, Jacob, Evgeniou, Theodoros, Toubia, Olivier, & Vert, J.-P. 2008. Eliciting consumer preferences using robust adaptive choice questionnaires. *Knowledge and Data Engineering, IEEE Transactions on*, **20**(2), 145–155. 00025.

Ahn, Jaemyung, de Weck, Olivier L., & Steele, Martin. 2014. Credibility Assessment of Models and Simulations Based on NASA's Models and Simulation Standard Using the Delphi Method: CREDIBILITY ASSESSMENT OF M&S USING THE DELPHI METHOD. *Systems Engineering*, **17**(2), 237–248. 00000.

Amazon. 2005. *Amazon mechanical turk*. http://www.mturk.com.

Andreassen, Erik, Clausen, Anders, Schevenels, Mattias, Lazarov, Boyan S., & Sigmund, Ole. 2011. Efficient topology optimization in MATLAB using 88 lines of code. *Structural and Multidisciplinary Optimization*, **43**(1), 1–16. 00146.

Antonsson, Erik K., & Cagan, Jonathan. 2005. *Formal engineering design synthesis*. Cambridge University Press. 00223.

Arrow, Kenneth J. 1951. *Social choice and individual values. 2nd*. Wiley, New York. 00166.

Austin-Breneman, Jesse, Yu, Bo Yang, & Yang, Maria C. 2014. Biased Information Passing Between Subsystems Over Time in Complex System Design. *Pages V007T07A023–V007T07A023 of: ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers. 00003.

Bachrach, Yoram, Graepel, Thore, Minka, Tom, & Guiver, John. 2012a. How to grade a test without knowing the answers - A Bayesian graphical model for adaptive crowdsourcing and aptitude testing. *In: Proceedings of the 29th International Conference on Machine Learning*.

Bachrach, Yoram, Graepel, Thore, Minka, Tom, & Guiver, John. 2012b (July). How To Grade a Test Without Knowing the Answers—A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. *Pages 1–9 of: Proceedings of the 29th International Conference on Machine Learning*.

Bao, Qifang, El Ferik, Sami, Shaukat, Mian Mobeen, & Yang, Maria C. 2014a (Aug.). An Investigation on the Inconsistency of Consumer Preferences: A Case Study of Residential Solar Panels. *In: Proceedings of the 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 00000.

Bao, Qifang, El Ferik, Sami, Shaukat, Mian Mobeen, & Yang, Maria C. 2014b (Aug.). An Investigation on the Inconsistency of Consumer Preferences: A Case Study of Residential Solar Panels. *In: Proceedings of the 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.

Barney, J. 1991. Firm Resources and Sustained Competitive Advantage. *Journal of Management*, **17**(1), 99–120. 37177.

Bayazit, Nigan. 2004. Investigating design: A review of forty years of design research. *Design issues*, **20**(1), 16–29. 00242.

Bayrak, Alparslan Emrah, Ren, Yi, & Papalambros, Panos Y. 2013a. Design of Hybrid-Electric Vehicle Architectures Using Auto-Generation of Feasible Driving Modes. *Pages V001T01A005–V001T01A005 of: ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers.

Bayrak, Alparslan Emrah, Ren, Yi, & Papalambros, Panos Y. 2013b. Design of Hybrid-Electric Vehicle Architectures Using Auto-Generation of Feasible Driving Modes. *Pages V001T01A005–V001T01A005 of: ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers.

Bengio, Yoshua. 2009. Learning deep architectures for AI. *Foundations and trends in Machine Learning*, **2**(1), 1–127.

Bengio, Yoshua, Goodfellow, Ian J, & Courville, Aaron. 2015. Deep Learning. *An MIT Press book in preparation. Draft chapters available at http://www. iro. umontreal. ca/ bengioy/dlbook*.

Bergstra, James, Breuleux, Olivier, Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Desjardins, Guillaume, Turian, Joseph, Warde-Farley, David, & Bengio, Yoshua. 2010 (June). Theano: a CPU and GPU Math Expression Compiler. *In: Proceedings of the Python for Scientific Computing Conference (SciPy).* Oral Presentation.

Berkovec, James, & Rust, John. 1985. A nested logit model of automobile holdings for one vehicle households. *Transportation Research Part B: Methodological*, **19**(4), 275–285.

Berlyne, Daniel E. 1971. *Aesthetics and Psychobiology.* East Norwalk, CT, US: Appleton-Century-Crofts. 02655.

Bjelland, Osvald M., & Wood, Robert Chapman. 2008. An Inside View of IBM's' Innovation Jam'. *MIT Sloan management review*, **50**(1), 32–40. 00179.

Björklund, Tua A. 2013. Initial mental representations of design problems: Differences between experts and novices. *Design Studies*, **34**(2), 135–160. 00012.

Bloch, Peter H. 1995. Seeking the Ideal Form: Product Design and Consumer Response. *Journal of Marketing*, **59**(3), 16. 00932.

Bloebaum, Christina, Collopy, Paul, & Hazelrigg, George A. 2012. NSF/NASA Workshop on the Design of Large-Scale Complex Engineered Systems - From Research to Product Realization. American Institute of Aeronautics and Astronautics. 00012.

Blohm, Ivo, Leimeister, Jan Marco, & Kremar, Helmut. 2013. Crowdsourcing: How to Benefit from (Too) Many Great Ideas. *MIS Quarterly Executive*, **12**(4), 189–200. 00000.

Bluntzer, J.-B., Ostrosi, E., & Sagot, J.-C. 2014. Styling of cars: is there a relationship between the style of cars and the culture identity of a specific country? *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, July. 00000.

Bodner, George M., & Guay, Roland B. 1997. The Purdue Visualization of Rotations Test. *The Chemical Educator*, **2**(4), 1–17. 00154.

Bommarito, M., Franzese R., Gong, A., & Page, S. 2011. Crowdsourcing design and evaluation analysis of DARPA's XC2V challenge. *University of Michigan Technical Report.*

Botsch, M., & Sorkine, O. 2008. On Linear Variational Surface Deformation Methods. *IEEE Transactions on Visualization and Computer Graphics*, **14**(1), 213–230. 00340.

Bradski, Gary, & Kaehler, Adrian. 2008. *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.".

Brin, Sergey, & Page, Lawrence. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, **30**(1), 107–117. 12673.

Budescu, David V, & Chen, Eva. 2014a. Identifying Expertise to Extract the Wisdom of Crowds. *Management Science*.

Budescu, David V., & Chen, Eva. 2014b. Identifying Expertise to Extract the Wisdom of the Crowds. *Management Science*. 00000.

Burnap, Alex, Hartley, Jeffrey, Pan, Yanxin, Gonzalez, Richard, & Papalambros, Panos Y. 2015a (Aug.). Balancing Design Freedom and Brand Recognition in the Evolution of Automotive Brand Character. *In: Proceedings of the 2015 International Design Engineering Technical Conferences*.

Burnap, Alex, Ren, Yi, Gerth, Richard, Papazoglou, Giannis, Gonzalez, Richard, & Papalambros, Panos Y. 2015b. When Crowdsourcing Fails: A Study of Expertise on Crowdsourced Design Evaluation. *Journal of Mechanical Design*, **137**(3), 031101.

Burnap, Alex, Ren, Yi, Gerth, Richard, Papazoglou, Giannis, Gonzalez, Richard, & Papalambros, Panos Y. 2015c. When Crowdsourcing Fails: A Study of Expertise on Crowdsourced Design Evaluation. *Journal of Mechanical Design*, **137**(3), 031101.

Burnap, Alex, Pan, Yanxin, Liu, Ye, Ren, Yi, Lee, Honglak, Gonzalez, Richard, & Papalambros, Panos Y. 2016. Improving Design Preference Prediction Accuracy With Feature Learning.

Caragiannis, Ioannis, Procaccia, Ariel D, & Shah, Nisarg. 2013. When do noisy votes reveal the truth? *Pages 143–160 of: Proceedings of the Fourteenth ACM Conference on Electronic Commerce*.

Celaschi, Flaviano, Celi, Manuela, & García, Laura Mata. 2011a. The extended value of design: an advanced design perspective. *Design Management Journal*, **6**(1), 6–15. 00015.

Celaschi, Flaviano, Celi, Manuela, & García, Laura Mata. 2011b. The extended value of design: An advanced design perspective. *Design Management Journal*, **6**(1), 6–15.

Chai, Chunlei, Cen, Fei, Ruan, Weiyu, Yang, Cheng, & Li, Hongting. 2015. Behavioral analysis of analogical reasoning in design: Differences among designers with different expertise levels. *Design Studies*, **36**(Jan.), 3–30. 00000.

Chandrasegaran, Senthil K., Ramani, Karthik, Sriram, Ram D., Horváth, Imré, Bernard, Alain, Harik, Ramy F., & Gao, Wei. 2013a. The evolution, challenges, and future of knowledge representation in product design systems. *Computer-Aided Design*, **45**(2), 204–228. 00132.

Chandrasegaran, Senthil K, Ramani, Karthik, Sriram, Ram D, Horváth, Imré, Bernard, Alain, Harik, Ramy F, & Gao, Wei. 2013b. The evolution, challenges, and future of knowledge representation in product design systems. *Computer-aided design*, **45**(2), 204–228.

Chen, Wei, Hoyle, Christopher, & Wassenaar, Henk Jan. 2013. *Decision-Based Design*. London: Springer London. 00011.

Chiu, Chao-Min, Liang, Ting-Peng, & Turban, Efraim. 2014. What can crowdsourcing do for decision support? *Decision Support Systems*, **65**(Sept.), 40–49.

Coates, Del. 2003. *Watches Tell More Than Time: Product Design, Information, and the Quest for Elegance*. McGraw-Hill London. 00125.

Condorcet, Marquis de. 1785. *Essay sur l'application de l'analyse de la probabilité des decisions: redues et pluralité des voix*. l'Imprimerie Royale. 00028.

Cooper, Robert G. 1990. Stage-gate systems: a new tool for managing new products. *Business horizons*, **33**(3), 44–54. 01260.

Crilly, Nathan. 2015. Fixation and creativity in concept development: The attitudes and practices of expert designers. *Design Studies*, **38**(May), 54–91. 00000.

Crilly, Nathan, Moultrie, James, & Clarkson, P.John. 2004a. Seeing things: consumer response to the visual domain in product design. *Design Studies*, **25**(6), 547–577. 00389.

Crilly, Nathan, Moultrie, James, & Clarkson, P.John. 2004b. Seeing things: consumer response to the visual domain in product design. *Design Studies*, **25**(6), 547–577.

Cross, Nigel. 2004a. Expertise in design: an overview. *Design Studies*, **25**(5), 427–441. 00537.

Cross, Nigel. 2004b. Expertise in design: an overview. *Design Studies*, **25**(5), 427–441.

Cross, Nigel. 2007. Forty years of design research. *Design Studies*, **28**(1), 1–4. 00179.

Dahlander, Linus, & Gann, David M. 2010. How open is innovation? *Research Policy*, **39**(6), 699–709. 01166.

Dalkey, Norman, & Helmer, Olaf. 1963. An experimental application of the Delphi method to the use of experts. *Management science*, **9**(3), 458–467. 02844.

Dauphin, Yann N, Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, Ganguli, Surya, & Bengio, Yoshua. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Pages 2933–2941 of: Advances in Neural Information Processing Systems*.

Davis-Stober, Clintin P, Budescu, David V, Dana, Jason, & Broomell, Stephen B. 2014. When is a crowd wise? *Decision*, **1**(2), 79.

Dawid, A. P., & Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, **28**(1), 20. 00429.

de Caritat, Marie Jean Antoine Nicolas, *et al.* 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale.

Della Penna, Nicolás, & Reid, Mark D. 2012. Crowd & Prejudice: An Impossibility Theorem for Crowd Labelling without a Gold Standard. *In: Proceedings of 2012 Collective Intelligence Conference*.

Diener, Kathleen, & Piller, Frank T. 2010. *The Market for Open Innovation: Increasing the efficiency and effectiveness of the innovation process*. RWTH Aachen University, Technology & Innovation Management Group.

Dinar, Mahmoud, Shah, Jami J., Cagan, Jonathan, Leifer, Larry, Linsey, Julie, Smith, Steven M., & Hernandez, Noe Vargas. 2015. Empirical Studies of Designer Thinking: Past, Present, and Future. *Journal of Mechanical Design*, **137**(2), 021101. 00002.

Dosovitskiy, Alexey, Springenberg, Jost Tobias, Tatarchenko, Maxim, & Brox, Thomas. 2014. Learning to Generate Chairs, Tables and Cars with Convolutional Networks. *arXiv:1411.5928 [cs]*, Nov.

Du, Ping, & MacDonald, Erin F. 2014. Eye-Tracking Data Predict Importance of Product Features and Saliency of Size Change. *Journal of Mechanical Design*, **136**(8), 081005.

Eckert, Claudia, & Stacey, Martin. 2000. Sources of inspiration: a language of design. *Design Studies*, **21**(5), 523–538.

Embretson, Susan E., & Reise, Steven P. 2013. *Item response theory*. Psychology Press.

Erickson, Lee B., Petrick, Irene, & Trauth, Eileen M. 2012. Organizational uses of the crowd: developing a framework for the study of crowdsourcing. *Pages 155–158 of: Proceedings of the 50th annual conference on Computers and People Research*. ACM. 00008.

Erickson, Lisa B. 2013. *Hanging with the right crowd: Crowdsourcing as a new business practice for innovation, productivity, knowledge capture, and marketing*. Ph.D. thesis, The Pennsylvania State University. 00000.

Ersal, Ilkin, Papalambros, Panos, Gonzalez, Richard, & Aitken, Thomas J. 2011. Modelling perceptions of craftsmanship in vehicle interior design. *Journal of Engineering Design*, **22**(2), 129–144. 00004.

Estelles-Arolas, E., & Gonzalez-Ladron-de Guevara, F. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science*, **38**(2), 189–200. 00536.

Estellés-Arolas, Enrique, & González-Ladrón-de Guevara, Fernando. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science*, **38**(2), 189–200.

Ester, Martin, Kriegel, Hans-Peter, Sander, J, & Xu, Xiaowei. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining*, **96**, 226–231.

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

Evgeniou, Theodoros, Pontil, Massimiliano, & Toubia, Olivier. 2007b. A Convex Optimization Approach to Modeling Consumer Heterogeneity in Conjoint Estimation. *Marketing Science*, **26**(6), 805–818. 00051.

Faas, Daniela, Bao, Qifang, Frey, Daniel D., & Yang, Maria C. 2014. The influence of immersion and presence in early stage engineering designing and building. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **28**(02), 139–151. 00000.

Field, Bruce W. 2007a. Visualization, Intuition, and Mathematics Metrics as Predictors of Undergraduate Engineering Design Performance. *Journal of Mechanical Design*, **129**(7), 735. 00004.

Field, Bruce W. 2007b. Visualization, Intuition, and Mathematics Metrics as Predictors of Undergraduate Engineering Design Performance. *Journal of Mechanical Design*, **129**(7), 735. 00004.

Flager, Forest, Gerber, David Jason, & Kallman, Ben. 2014. Measuring the impact of scale and coupling on solution quality for building design problems. *Design Studies*, **35**(2), 180–199. 00008.

Francis, Paul, Golden, Michael, & Woods, William. 2010. *Defense Acquisitions: Managing Risk to Achieve Better Outcomes*. Tech. rept. DTIC Document. 00001.

Fu, Katherine, Chan, Joel, Schunn, Christian, Cagan, Jonathan, & Kotovsky, Kenneth. 2013a. Expert representation of design repository space: A comparison to and validation of algorithmic output. *Design Studies*, **34**(6), 729–762. 00002.

Fu, Katherine, Chan, Joel, Cagan, Jonathan, Kotovsky, Kenneth, Schunn, Christian, & Wood, Kristin. 2013b. The meaning of "near" and "far": the impact of structuring design databases and the effect of distance of analogy on design output. *Journal of Mechanical Design*, **135**(2), 021007. 00047.

Fu, Katherine, Murphy, Jeremy, Yang, Maria, Otto, Kevin, Jensen, Dan, & Wood, Kristin. 2015. Design-by-analogy: experimental evaluation of a functional analogy search methodology for concept generation improvement. *Research in Engineering Design*, **26**(1), 77–95.

Fu, Luoting, & Kara, Levent Burak. Deciphering the Influence of Product Shape on Consumer Judgments Through Geometric Abstraction.

Fuge, Mark. 2015. A Scalpel not a Sword: On the Role of Statistical Tests in Design Cognition. *Pages 1–11 of: ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers.

Fuge, Mark, Stroud, Josh, & Agogino, Alice. 2013. Automatically Inferring Metrics for Design Creativity. *Page V005T06A010 of: Proceedings of the ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Chicago, IL, USA: American Society of Mechanical Engineers.

Gelfand, Alan E, & Smith, Adrian FM. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**(410), 398–409.

Gero, John S., & Maher, Mary Lou. 2013. *Modeling Creativity and Knowledge-Based Creative Design*. Psychology Press.

Gerth, Richard J, Burnap, Alex, & Papalambros, Panos. 2012. Crowdsourcing: A Primer and Its Implications for Systems Engineering. *In: 2012 NDIA Ground Vehicle Systems Engineering and Technology Symposium*.

Goldschmidt, Gabriela. 1997. Capturing indeterminism: representation in the design problem space. *Design Studies*, **18**(4), 441–455.

Gonzalez, Richard, & Nelson, Thomas O. 1996. Measuring ordinal association in situations that contain tied scores. *Psychological bulletin*, **119**(1), 159. 00067.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, & Bengio, Yoshua. 2014. Generative adversarial nets. *Pages 2672–2680 of: Advances in Neural Information Processing Systems*. 00017.

Grace, Kazjon, Maher, Mary Lou, Fisher, Douglas, & Brady, Katherine. 2014. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation*, 1–23.

Green, Paul E., Carroll, J. Douglas, & Goldberg, Stephen M. 1981. A General Approach to Product Design Optimization Via Conjoint Analysis. *Journal of Marketing*, **45**(3), 17. 00374.

Gurnani, Ashwin, & Lewis, Kemper. 2008. Collaborative, decentralized engineering design at the edge of rationality. *Journal of Mechanical Design*, **130**(12), 121101.

Haario, Heikki, Saksman, Eero, & Tamminen, Johanna. 2001. An adaptive Metropolis algorithm. *Bernoulli*, **7**(2), 223–242.

Hartley, Jeffrey. 1996a. *Brands Through the Lens of Style*. 00000.

Hartley, Jeffrey. 1996b. Brands Through the Lens of Style. San Diego, California: Quest and Associates.

Hauser, John, Tellis, Gerard J., & Griffin, Abbie. 2006. Research on Innovation: A Review and Agenda for Marketing Science. *Marketing Science*, **25**(6), 687–717. 00792.

Hazelrigg, George A. 1998. A framework for decision-based engineering design. *Journal of mechanical design*, **120**(4), 653–658. 00462.

Hegarty, Mary. 2004. Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, **8**(6), 280–285. 00175.

Hekkert, Paul, Snelders, Dirk, & Wieringen, Piet CW. 2003. 'Most advanced, yet acceptable': typicality and novelty as joint predictors of aesthetic preference in industrial design. *British Journal of Psychology*, **94**(1), 111–124. 00209.

Herbrich, Ralf, Graepel, Thore, Bollmann-Sdorra, Peter, & Obermayer, Klaus. 1998. Learning preference relations for information retrieval. *Pages 80–84 of: ICML-98 Workshop: text categorization and machine learning*. 00098.

Ho, Chun-Heng. 2001a. Some phenomena of problem decomposition strategy for design thinking: differences between novices and experts. *Design Studies*, **22**(1), 27–45. 00118.

Ho, Chun-Heng. 2001b. Some phenomena of problem decomposition strategy for design thinking: differences between novices and experts. *Design Studies*, **22**(1), 27–45.

Hong, Lu, & Page, Scott E. 2004a. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(46), 16385–16389. 00314.

Hong, Lu, & Page, Scott E. 2004b. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(46), 16385–16389.

Hooshmand, Amir, & Campbell, Matthew I. 2014. Layout synthesis of fluid channels using generative graph grammars. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **28**(03), 239–257. 00001.

Howe, Jeff. 2006. The rise of crowdsourcing. *Wired magazine*, **14**(6), 1–4. 02815.

Huizingh, Eelko K.R.E. 2011. Open innovation: State of the art and future perspectives. *Technovation*, **31**(1), 2–9. 00742.

Hüllermeier, Eyke, Fürnkranz, Johannes, Cheng, Weiwei, & Brinker, Klaus. 2008. Label ranking by learning pairwise preferences. *Artificial Intelligence*, **172**(16–17), 1897–1916. 00285.

Ind, Nicholas, & Watt, Cameron. 2006. Brands and breakthroughs: How brands help focus creative decision making. *The Journal of Brand Management*, **13**(4-5), 330–338. 00011.

Ipeirotis, Panagiotis G, & Paritosh, Praveen K. 2011. Managing crowdsourced human computation: a tutorial. *Pages 287–288 of: Proceedings of the 20th International World Wide Web Conference Companion*.

Kang, Sung Woo, & Tucker, Conrad S. 2015. Automated Concept Generation Based on Function-Form Synthesis. *Pages V02AT03A008–V02AT03A008 of: ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers.

Kavakli, Manolya, & Gero, John S. 2001a. Sketching as mental imagery processing. *Design Studies*, **22**(4), 347–364. 00127.

Kavakli, Manolya, & Gero, John S. 2001b. Sketching as mental imagery processing. *Design Studies*, **22**(4), 347–364.

Keller, Kevin Lane. 2003. Brand Synthesis: The Multidimensionality of Brand Knowledge. *Journal of Consumer Research*, **29**(4), 595–600. 01105.

Kim, Juho, Zhang, Haoqi, André, Paul, Chilton, Lydia B, Mackay, Wendy, Beaudouin-Lafon, Michel, Miller, Robert C, & Dow, Steven P. 2013. Cobi: A community-informed conference scheduling tool. *Pages 173–182 of: Proceedings of the 26th Annual ACM symposium on User Interface Software and Technology*.

Kingma, Diederik P., & Adam, Jimmy Ba. 2015. Adam: A method for stochastic optimization. *In: International Conference on Learning Representation*.

Kingma, Diederik P., & Welling, Max. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, Diederik P., Mohamed, Shakir, Rezende, Danilo Jimenez, & Welling, Max. 2014. Semi-supervised learning with deep generative models. *Pages 3581–3589 of: Advances in Neural Information Processing Systems*.

Kittur, Aniket, Chi, Ed H, & Suh, Bongwon. 2008. Crowdsourcing user studies with Mechanical Turk. *Pages 453–456 of: Proceedings of the SIGCHI conference on human factors in computing systems*.

Klein, Mark, Sayama, Hiroki, Faratin, Peyman, & Bar-Yam, Yaneer. 2006. The dynamics of collaborative design: insights from complex systems and negotiation research. *Pages 158–174 of: Complex Engineered Systems*. Springer. 00014.

Kókai, István, Finger, Jörg, Smith, Randall C., Pawlicki, Richard, & Vetter, Thomas. 2007a (Aug.). Example-Based Conceptual Styling Framework for Automotive Shapes. *Pages 37–44 of: Proceedings of the 4th Eurographics Workshop on Sketch-Based Interfaces and Modeling*. 00014.

Kókai, István, Finger, Jörg, Smith, Randall C, Pawlicki, Richard, & Vetter, Thomas. 2007b. Example-based conceptual styling framework for automotive shapes. *Pages 37–44 of: Proceedings of the 4th Eurographics workshop on Sketch-based interfaces and modeling*. ACM.

Kozhevnikov, Maria, Motes, Michael A., & Hegarty, Mary. 2007. Spatial visualization in physics problem solving. *Cognitive Science*, **31**(4), 549–579. 00093.

Kreuzbauer, Robert, & Malter, Alan J. 2005. Embodied Cognition and New Product Design: Changing Product Form to Influence Brand Categorization. *Journal of Product Innovation Management*, **22**(2), 165–176. 00000.

Krishnan, Viswanathan, & Ulrich, Karl T. 2001. Product development decisions: A review of the literature. *Management science*, **47**(1), 1–21. 01166.

Krizhevsky, Alex, Sutskever, Ilya, & Hinton, Geoffrey E. 2012. Imagenet classification with deep convolutional neural networks. *Pages 1097–1105 of: Advances in neural information processing systems*.

Kruger, Justin, Endriss, Ulle, Fernández, Raquel, & Qing, Ciyang. 2014. Axiomatic analysis of aggregation methods for collective annotation. *Pages 1185–1192 of: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*.

Kudrowitz, Barry Matthew, & Wallace, David. 2013. Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design*, **24**(2), 120–139.

Lakshminarayanan, Balaji, & Teh, Yee Whye. 2013. Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. *arXiv preprint arXiv:1305.0015*.

Lau, Kimberly, Beckman, Sara L., & Agogino, Alice M. 2012. Diversity in design teams: An investigation of learning styles and their impact on team performance and innovation. *International Journal of Engineering Education*, **28**(2), 293. 00002.

Lee, Hau, & Whang, Seungjin. 1999. Decentralized multi-echelon supply chains: Incentives and information. *Management Science*, **45**(5), 633–640. 00587.

Lee, Honglak, Ekanadham, Chaitanya, & Ng, Andrew Y. 2008. Sparse Deep Belief Net Model for Visual Area V2. *Advances in Neural Information Processing Systems 20*, 873–880.

Lee, Honglak, Grosse, Roger, Ranganath, Rajesh, & Ng, Andrew Y. 2011. Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks. *Communications of the Association for Computing Machinery*, **54**(10), 95–103.

Lewis, Kemper E, Chen, Wei, & Schmidt, Linda C. 2006. *Decision making in engineering design*. American Society of Mechanical Engineers.

Li, Hongwei, & Yu, Bin. 2014. Error Rate Bounds and Iterative Weighted Majority Voting for Crowdsourcing. *arXiv preprint arXiv:1411.4086*. 00000.

Linsey, Julie Stahmer. 2007. *Design-by-analogy and representation in innovative engineering concept generation*. ProQuest.

Linstone, Harold A., Turoff, Murray, & others. 1975. *The Delphi method: Techniques and applications*. Vol. 29. Addison-Wesley Reading, MA. 05520.

Liu, Qiang, Peng, Jian, & Ihler, Alexander T. 2012. Variational Inference for Crowdsourcing. *Advances in Neural Information Processing Systems*, 701–709.

Lord, Frederic. 1952. A theory of test scores. *Psychometric monographs*. 00699.

Lord, Frederic M. 1980. *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Louizos, Christos, Swersky, Kevin, Li, Yujia, Zemel, Richard, & Welling, Max. 2015. The Variational Fair Autoencoder. *arXiv:1511.00830 [cs, stat]*, Nov.

Louridas, Panagiotis. 1999. Design as bricolage: anthropology meets design thinking. *Design Studies*, **20**(6), 517–535. 00111.

Luce, R. Duncan. 1959. On the possible psychophysical laws. *Psychological review*, **66**(2), 81. 00363.

Lugo, José E., Schmiedeler, James P., Batill, Stephen M., & Carlson, Laura. 2015. Quantification of Classical Gestalt Principles in Two-Dimensional Product Representations. *Journal of Mechanical Design*, **137**(9), 094502. 00000.

Ma, Jungmok, Kwak, Minjung, & Kim, Harrison M. 2014. Demand Trend Mining for Predictive Life Cycle Design. *Journal of Cleaner Production*, **68**(Apr.), 189–199. 00005.

MacDonald, Erin F., Gonzalez, Richard, & Papalambros, Panos Y. 2009. Preference Inconsistency in Multidisciplinary Design Decision Making. *Journal of Mechanical Design*, **131**(3), 031009. 00053.

Mannering, Fred, Winston, Clifford, Griliches, Zvi, & Schmalensee, Richard. 1991. Brand Loyalty and the Decline of American Automobile Firms. *Brookings Papers on Economic Activity. Microeconomics*, **1991**(Jan.), 67–114. 00102.

Manoogian II, John. 2013 (June). *Vehicle Design Process used at General Motors*. 00000.

Martindale, Colin. 1990. *The Clockwork Muse: The Predictability of Artistic Change*. New York, NY: Basic Books. 00368.

McCormack, Jay P, Cagan, Jonathan, & Vogel, Craig M. 2004a. Speaking the Buick language: capturing, understanding, and exploring brand identity with shape grammars. *Design Studies*, **25**(1), 1–29.

McCormack, Jay P, Cagan, Jonathan, & Vogel, Craig M. 2004b. Speaking the Buick language: capturing, understanding, and exploring brand identity with shape grammars. *Design Studies*, **25**(1), 1–29. 00187.

McFadden, Daniel, & others. 1973. Conditional logit analysis of qualitative choice behavior. 12600.

McFadden, Daniel, & Train, Kenneth. 2000. Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, **15**(5), 447–470.

McGowan, Anna-Maria Rivas, Daly, Shanna, Baker, Wayne, Papalambros, Panos, & Seifert, Colleen. 2013. A Socio-Technical Perspective on Interdisciplinary Interactions During the Development of Complex Engineered Systems. *Procedia Computer Science*, **16**, 1142–1151. 00003.

McKenna, Ann F., & Agogino, Alice M. 2004. Supporting Mechanical Reasoning with a Representationally-Rich Learning Environment. *Journal of Engineering Education*, **93**(2), 97–104. 00016.

McWilliam, Gil, & Dumas, Angela. 1997. Using metaphors in new brand design. *Journal of Marketing Management*, **13**(4), 265–284. 00026.

Michalek, J.J., Feinberg, F.M., & Papalambros, P.Y. 2005. Linking marketing and engineering product design decisions via analytical target cascading. *Journal of Product Innovation Management*, **22**(1), 42–62.

Milburn, C, & Childs, Peter R. N. 2001. The Styling Process. *Pages 275–295 of: Total Vehicle Technology: Challenging Current Thinking*. John Wiley & Sons. 00001.

Miller, Nolan, Resnick, Paul, & Zeckhauser, Richard. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science*, **51**(9), 1359–1373.

Miller, Scarlett R., Bailey, Brian P., & Kirlik, Alex. 2014. Exploring the Utility of Bayesian Truth Serum for Assessing Design Knowledge. *Human–Computer Interaction*, **29**(5-6), 487–515. 00003.

Moreno, Diana P., Hernández, Alberto A., Yang, Maria C., Otto, Kevin N., Hölttä-Otto, Katja, Linsey, Julie S., Wood, Kristin L., & Linden, Adriana. 2014. Fundamental studies in Design-by-Analogy: A focus on domain-knowledge experts and

applications to transactional design problems. *Design Studies*, **35**(3), 232–272. 00010.

Moulson, Tom, & Sproles, George. 2000. Styling Strategy. *Business Horizons*, **43**(5), 45–52. 00032.

Mukherjee, Arpan, Zhang, Yunbo, & Rai, Rahul. 2014. Probabilistic Design Miming. *Pages V007T07A011–V007T07A011 of: ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers.

Mulder-Nijkamp, Maaike, & Eggink, Wouter. 2013. Brand Value by Design: The Use of Three Levels of Recognition in Design. *In: Proceedings of the 5th International Congress of International Association of Societies of Design Research*. 00001.

Murugappan, Sundar, Piya, Cecil, Ramani, Karthik, & others. 2013. Handy-potter: Rapid exploration of rotationally symmetric shapes through natural hand motions. *Journal of Computing and Information Science in Engineering*, **13**(2), 021008.

Norman, Donald A. 2004. *Emotional Design: Why We Love (or Hate) Everyday Things*. New York, NY: Basic books. 03784.

Norman, Donald A. 2007. *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books. 00000.

Norman, Donald A., Ortony, Andrew, & Russell, Daniel M. 2003. Affect and machine design: Lessons for the development of autonomous machines. *IBM Systems Journal*, **42**(1), 38–44. 00133.

Nunnally, Jum, & Bernstein, Ira. 2010. *Psychometric Theory 3E*. McGraw-Hill series in psychology. McGraw-Hill Education.

Oberhauser, Matthias, Sartorius, Sky, Gmeiner, Thomas, & Shea, Kristina. 2015. Computational Design Synthesis of Aircraft Configurations with Shape Grammars. *Pages 21–39 of: Design Computing and Cognition'14*. Springer.

Oehmen, J., & Seering, W. 2011. Risk-Driven Design Processes: Balancing Efficiency with Resilience in Product Design. *Pages 47–54 of:* Birkhofer, Herbert (ed), *The Future of Design Methodology*. London: Springer London. 00000.

Oravecz, Zita, Anders, Royce, & Batchelder, William H. 2013. Hierarchical Bayesian modeling for test theory without an answer key. *Psychometrika*, 1–24.

Orbay, Gunay, Fu, Luoting, & Kara, Levent Burak. 2015a. Deciphering the Influence of Product Shape on Consumer Judgments Through Geometric Abstraction. *Journal of Mechanical Design*, **137**(8), 081103. 00000.

Orbay, Gunay, Fu, Luoting, & Kara, Levent Burak. 2015b. Deciphering the Influence of Product Shape on Consumer Judgments Through Geometric Abstraction. *Journal of Mechanical Design*, Mar. 00000.

Orbay, Gunay, Fu, Luoting, & Kara, Levent Burak. 2015c. Deciphering the Influence of Product Shape on Consumer Judgments Through Geometric Abstraction. *Journal of Mechanical Design*, **137**(8), 081103.

Orsborn, Seth, & Cagan, Jonathan. 2009. Multiagent Shape Grammar Implementation: Automatically Generating Form Concepts According to a Preference Function. *Journal of Mechanical Design*, **131**(12), 121007. 00020.

Orsborn, Seth, Cagan, Jonathan, Pawlicki, Richard, & Smith, Randall C. 2006a. Creating cross-over vehicles: Defining and combining vehicle classes using shape grammars. *AIE EDAM: Artificial Intelligence for Engineering Design, Analysis, and Manufacturing*, **20**(03), 217–246. 00046.

Orsborn, Seth, Cagan, Jonathan, Pawlicki, Richard, & Smith, Randall C. 2006b. Creating cross-over vehicles: Defining and combining vehicle classes using shape grammars. *AIE EDAM: Artificial Intelligence for Engineering Design, Analysis, and Manufacturing*, **20**(03), 217–246.

Orsborn, Seth, Cagan, Jonathan, & Boatwright, Peter. 2009a. Quantifying Aesthetic Form Preference in a Utility Function. *Journal of Mechanical Design*, **131**(6), 061001. 00052.

Orsborn, Seth, Cagan, Jonathan, & Boatwright, Peter. 2009b. Quantifying Aesthetic Form Preference in a Utility Function. *Journal of Mechanical Design*, **131**(6).

Ozkan, Ozgu, & Dogan, Fehmi. 2013. Cognitive strategies of analogical reasoning in design: Differences between expert and novice designers. *Design Studies*, **34**(2), 161–192. 00020.

Pan, Yanxin, Burnap, Alex, Liu, Ye, Lee, Honglak, Gonzalez, Richard, & Papalambros, Panos. 2016 (May). A Quantitative Model for Identifying Regions of Design Visual Attraction and Application to Automobile Styling. *In: Proceedings of the 2016 Internation Design Conference*.

Panchal, Jitesh. 2015a. Using Crowds in Engineering DesignTowards a Holistic Framework. *Pages 1–10 of: Proceedings of the 2015 International Conference on Engineering Design*. Design Society.

Panchal, Jitesh H. 2015b. Using Crowds in Engineering Design–Towards a Holistic Framework. 00000.

Panchal, Jitesh H. 2015c. Using Crowds in Engineering DesignTowards a Holistic Framework. 00000.

Papalambros, Panos Y. 2002 (July). An Enterprize Context for Design Optimization. *In: Proceedings of the ESDA 2002 6th Biennial Conference on Engineering Systems Design and Analysis*. 00004.

Papalambros, Panos Y., & Chirehdast, Mehran. 1990. An Integrated Environment for Structural Configuration Design. *Journal of Engineering Design*, **1**(1), 73–96. 00081.

Papalambros, Panos Y., & Shea, Kristina. 2005. Creating Structural Configurations. *Pages 93–125 of:* Antonsson, Erik K, & Cagan, Jonathan (eds), *Formal engineering design synthesis*. Cambridge University Press.

Papalambros, Panos Y., & Wilde, Douglass J. 2000. *Principles of Optimal Design: Modeling and Computation*. Cambridge University Press. 01124.

Patil, Anand, Huard, David, & Fonnesbeck, Christopher J. 2010. PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, **35**(4), 1.

Peisl, Thomas, Selen, Willem, Raeside, Robert, & Albera, Tatiana. 2014. Predictive Crowding as a Concept to Support the Assessment of Disruptive Ideas: A Conceptual Framework. *Journal of New Business Ideas & Trends*, **12**(2), 1–13.

Perez Mata, Marta, Ahmed-Kristensen, Saeema, & Shea, Kristina. 2015. Spatial Grammar for Design Synthesis Targeting Perceptions: Case Study on Beauty. Aug., V01AT02A013.

Person, Oscar, & Snelders, Dirk. 2010. Brand styles in commercial design. *Design Issues*, **26**(1), 82–94. 00009.

Person, Oscar, Snelders, Dirk, Karjalainen, Toni-Matti, & Schoormans, Jan. 2007. Complementing Intuition: Insights on Styling as a Strategic Tool. *Journal of Marketing Management*, **23**(9-10), 901–916. 00017.

Person, Oscar, Schoormans, Jan, Snelders, Dirk, & Karjalainen, Toni-Matti. 2008. Should new products look similar or different? The influence of the market environment on strategic product styling. *Design Studies*, **29**(1), 30–48. 00032.

Petiot, Jean-Fran\ccois, Salvo, Ccile, Hossoy, Ilkin, & Papalambros, Panos Y. 2009. A cross-cultural study of users' craftsmanship perceptions in vehicle interior design. *International Journal of Product Development*, **7**(1), 28–46.

Petiot, Jean-François, & Dagher, Antoine. 2010. Preference-Oriented Form Design: Application to Cars' Headlights. *International Journal on Interactive Design and Manufacturing*, **5**(1), 17–27. 00000.

Pilz, Dennis, & Gewald, Heiko. 2013. Does money matter? Motivational factors for participation in paid-and non-profit-crowdsourcing communities. *Wirtschaftsinformatik Proceedings 2013*. 00007.

Plackett, Robin L. 1975. The analysis of permutations. *Applied Statistics*, 193–202. 00270.

Poirson, Emilie, Petiot, Jean-François, Boivin, Ludivine, & Blumenthal, David. 2013a. Eliciting user perceptions using assessment tests based on an interactive genetic algorithm. *Journal of Mechanical Design*, **135**(3), 031004. 00004.

Poirson, Emilie, Petiot, Jean-Franois, Boivin, Ludivine, & Blumenthal, David. 2013b. Eliciting user perceptions using assessment tests based on an interactive genetic algorithm. *Journal of Mechanical Design*, **135**(3), 031004.

Powell, Michael JD. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, **7**(2), 155–162.

Prelec, D. 2004a. A Bayesian Truth Serum for Subjective Data. *Science*, **306**(5695), 462–466. 00156.

Prelec, Dražen. 2004b. A Bayesian truth serum for subjective data. *Science*, **306**(5695), 462–466.

Prelec, Drazen, & Seung, H. Sebastian. 2007. *An algorithm that finds truth even if most people are wrong.* Tech. rept. Working paper. 00011.

Prelec, Drazen, Seung, H Sebastian, & McCoy, John. 2013. *Finding truth even if the crowd is wrong.* Tech. rept. Working paper, MIT.

Pugliese, Michael J., & Cagan, Jonathan. 2002a. Capturing a rebel: modeling the Harley-Davidson brand through a motorcycle shape grammar. *Research in Engineering Design*, **13**(3), 139–156. 00124.

Pugliese, Michael J., & Cagan, Jonathan. 2002b. Capturing a rebel: modeling the Harley-Davidson brand through a motorcycle shape grammar. *Research in Engineering Design*, **13**(3), 139–156.

Ramani, Karthik, Lee Jr, Kevin, Jasti, Raja, & others. 2014. zPots: a virtual pottery experience with spatial interactions using the leap motion device. *Pages 371–374 of: CHI'14 Extended Abstracts on Human Factors in Computing Systems.* ACM. 00001.

Ramanujan, Devarajan, Vinayak, Yash Nawal, Reid, Tahira, & Ramani, Karthik. Informing early design via crowd-based co-creation. 00000.

Ranawat, Arjun, Tuteja, Sumit, & Höltta–Otto, Katja. 2012. Contribution of Visual Design Elements to the Perceived Product Family Look. *Journal of Design Research*, **10**(3), 189–205. 00001.

Ranscombe, Charlie, Hicks, Ben, Mullineux, Glen, & Singh, Baljinder. 2012. Visually decomposing vehicle images: Exploring the influence of different aesthetic features on consumer perception of brand. *Design Studies*, **33**(4), 319–341. 00008.

Rasch, Georg. 1960/1980. Probabilistic models for some intelligence and achievement tests, expanded edition (1980) with foreword and afterword by B.D. Wright. *Copenhagen, Denmark: Danish Institute for Educational Research.*

Rasch, Georg. 1966. An item analysis which takes individual differences into account. *British journal of mathematical and statistical psychology*, **19**(1), 49–57. 00369.

Rasoulifar, Golnoosh, Prudhomme, Guy, & Eckert, Claudia. 2015. Communicating Consumer Needs in the Design Process of Branded Products. *Journal of Mechanical Design*, Mar. 00000.

Raykar, Vikas C., Yu, Shipeng, Zhao, Linda H., Jerebko, Anna, Florin, Charles, Valadez, Gerardo Hermosillo, Bogoni, Luca, & Moy, Linda. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. ACM. 00126.

Raykar, Vikas C, Yu, Shipeng, Zhao, Linda H, Valadez, Gerardo Hermosillo, Florin, Charles, Bogoni, Luca, & Moy, Linda. 2010. Learning from crowds. *The Journal of Machine Learning Research*, **11**, 1297–1322.

Reid, Tahira N., Gonzalez, Richard D., & Papalambros, Panos Y. 2010a. Quantification of Perceived Environmental Friendliness for Vehicle Silhouette Design. *Journal of Mechanical Design*, **132**(10), 101010.

Reid, Tahira N, Gonzalez, Richard D, & Papalambros, Panos Y. 2010b. Quantification of Perceived Environmental Friendliness for Vehicle Silhouette Design. *Journal of Mechanical Design*, **132**, 101010.

Reid, Tahira N, Frischknecht, Bart D, & Papalambros, Panos Y. 2012. Perceptual Attributes in Product Design: Fuel Economy and Silhouette-Based Perceived Environmental Friendliness Tradeoffs in Automotive Vehicle Design. *Journal of Mechanical Design*, **134**, 041006.

Reid, Tahira N., MacDonald, Erin F., & Du, Ping. 2013a. Impact of Product Design Representation on Customer Judgment. *Journal of Mechanical Design*, **135**(9), 091008. 00018.

Reid, Tahira N., MacDonald, Erin F., & Du, Ping. 2013b. Impact of Product Design Representation on Customer Judgment. *Journal of Mechanical Design*, **135**(9), 091008.

Ren, Yi, & Papalambros, Panos Y. 2012a. On Design Preference Elicitation With Crowd Implicit Feedback. *Pages 541–551 of: Proceedings of the 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.*

Ren, Yi, & Papalambros, Panos Y. 2012b (Aug.). On Design Preference Elicitation with Crowd Implicit Feedback. *In: Proceedings of the 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.* 00002.

Ren, Yi, Burnap, Alex, & Papalambros, Panos. 2013a (Aug.). Quantification of Perceptual Design Attributes Using a Crowd. *In: Proceedings of the 19th International Conference on Engineering Design*.

Ren, Yi, Burnap, Alex, Papalambros, Panos, *et al.* 2013b. Quantification of perceptual design attributes using a crowd. *In: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol. 6: Design Information and Knowledge*.

Ren, Yi, Bayrak, Emrah, & Papalambros, Panos Y. 2015a (Aug.). Ecoracer: Optimal Design and Control of Electric Vehicles Using Human Game Players. *In: Proceedings of the ASME 2015 International Design Engineering Technical Conferences*. 00000.

Ren, Yi, Bayrak, Emrah, & Papalambros, Panos Y. 2015b. Ecoracer: Optimal Design and Control of Electric Vehicles Using Human Game Players. *In: Proceedings of the ASME 2015 International Design Engineering Technical Conferences*.

Rosenman, Michael A., & Gero, John S. 1993. Creativity in design using a design prototype approach. *Modeling creativity and knowledge-based creative design*, 111–138.

Ross, Sheldon M. 1996. *Stochastic processes*. Vol. 2. John Wiley & Sons New York. 06318.

Rzeszotarski, Jeffrey M, & Kittur, Aniket. 2011. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. *Pages 13–22 of: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*.

Salakhutdinov, Ruslan, Mnih, Andriy, & Hinton, Geoffrey. 2007. Restricted Boltzmann machines for collaborative filtering. *Proceedings of the 24th International Conference on Machine Learning*, 791–798.

Schmidhuber, Jürgen. 2015. Deep Learning in Neural Networks: An Overview. *Neural Networks*, **61**, 85–117.

Schmitt, Bernd. 2012. The consumer psychology of brands. *Journal of Consumer Psychology*, **22**(1), 7–17. 00042.

Schramm, U, Thomas, HL, Zhou, M, & Voth, B. 1999. Topology optimization with Altair OptiStruct. *In: Proceedings of the Optimization in Industry II Conference*.

Shah, Jami J., Woodward, Jay, & Smith, Steven M. 2013. Applied Tests of Design Skills—Part II: Visual Thinking. *Journal of Mechanical Design*, **135**(7), 071004. 00004.

Shankar, S. Sree, & Rai, Rahul. 2014. Human factors study on the usage of BCI headset for 3D CAD modeling. *Computer-Aided Design*, **54**, 51–55. 00002.

Sheng, Victor S, Provost, Foster, & Ipeirotis, Panagiotis G. 2008a. Get another label? improving data quality and data mining using multiple, noisy labelers. *Pages 614–622 of: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Las Vegas, NV, USA: Association for Computing Machinery.

Sheng, Victor S., Provost, Foster, & Ipeirotis, Panagiotis G. 2008b. Get another label? improving data quality and data mining using multiple, noisy labelers. *Pages 614–622 of: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 00415.

Sheshadri, Aashish, & Lease, Matthew. 2013a. SQUARE: A Benchmark for Research on Computing Crowd Consensus. *In: Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing*.

Sheshadri, Aashish, & Lease, Matthew. 2013b (Nov.). SQUARE: A Benchmark for Research on Computing Crowd Consensus. *In: Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*.

Simon, Herbert A. 1956. Rational choice and the structure of the environment. *Psychological review*, **63**(2), 129. 03836.

Simon, Herbert A. 1969. The sciences of the artificial. *Cambridge, MA*. 00590.

Simon, Herbert A. 1996. *The sciences of the artificial*. MIT press.

Simonyan, Karen, Vedaldi, Andrea, & Zisserman, Andrew. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]*, Dec.

Simpson, Timothy W., Rosen, David, Allen, Janet K., & Mistree, Farrokh. 1998. Metrics for assessing design freedom and information certainty in the early stages of design. *Journal of Mechanical Design*, **120**(4), 628–635. 00073.

Smith, Joshua R. 1991. Designing Biomorphs with an Interactive Genetic Algorithm. *Pages 535–538 of: ICGA*.

Snow, Rion, O'Connor, Brendan, Jurafsky, Daniel, & Ng, Andrew Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Pages 254–263 of: Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Srinivasan, Raji, Lilien, Gary L., & Rangaswamy, Arvind. 2006. The Emergence of Dominant Designs. *Journal of Marketing*, **70**(2), 1–17. 00091.

Stone, Thomas, & Choi, Seung-Kyum. 2013. Extracting Consumer Preference From User-Generated Content Sources Using Classification. *Page V03AT03A031 of: Proceedings of the ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.

Sylcott, Brian, Michalek, Jeremy J, & Cagan, Jonathan. 2013a. Towards understanding the role of interaction effects in visual conjoint analysis. *Pages V03AT03A012–V03AT03A012 of: ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers.

Sylcott, Brian, Cagan, Jonathan, & Tabibnia, Golnaz. 2013b. Understanding consumer tradeoffs between form and function through metaconjoint and cognitive neuroscience analyses. *Journal of Mechanical Design*, **135**(10), 101002. 00014.

Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian, & Fergus, Rob. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Talke, Katrin, Salomo, Sören, Wieringa, Jaap E., & Lutz, Antje. 2009. What about design newness? Investigating the relevance of a neglected dimension of product innovativeness. *Journal of Product Innovation Management*, **26**(6), 601–615. 00071.

Tang, Wei, & Lease, Matthew. 2011. Semi-supervised consensus labeling for crowdsourcing. *Special Interest Group on Information Retrieval 2011 Workshop on Crowdsourcing for Information Retrieval*.

Telenko, Cassandra, & Wood, Kristin. INNOVATIVE AND SUSTAINABLE DESIGN: PERCEPTIONS OF EXPERTS. 00000.

Telenko, Cassandra, Sosa, Ricardo, & Wood, Kristin L. 2016. Changing Conversations and Perceptions: The Research and Practice of Design Science. *Pages 281–309 of: Impact of Design Research on Industrial Practice*. Springer.

Theis, Lucas, Oord, Aron van den, & Bethge, Matthias. 2015. A note on the evaluation of generative models. *arXiv:1511.01844 [cs, stat]*, Nov.

Thurstone, Louis L. 1927. A law of comparative judgment. *Psychological review*, **34**(4), 273. 04444.

Tiwari, Santosh, Dong, Hong, Fadel, Georges, Fenyes, Peter, & Kloess, Artemis. 2014. A physically-based shape morphing algorithm for packing and layout applications. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 1–13. 00000.

Toh, Christine A., & Miller, Scarlett R. 2014a. The Impact of Example Modality and Physical Interactions on Design Creativity. *Journal of Mechanical Design*. 00000.

Toh, Christine A., & Miller, Scarlett R. 2014b. The Impact of Example Modality and Physical Interactions on Design Creativity. *Journal of Mechanical Design*.

Tovares, Noah, Boatwright, Peter, & Cagan, Jonathan. 2014a. Experiential Conjoint Analysis: An Experience-Based Method for Eliciting, Capturing, and Modeling Consumer Preference. *Journal of Mechanical Design*, **136**(10), 101404. 00000.

Tovares, Noah, Boatwright, Peter, & Cagan, Jonathan. 2014b. Experiential Conjoint Analysis: An Experience-Based Method for Eliciting, Capturing, and Modeling Consumer Preference. *Journal of Mechanical Design*, **136**(10), 101404.

Tuarob, Suppawong, & Tucker, Conrad S. 2013. Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data. *Page V02BT02A012 of: Proceedings of the ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Portland, OR, USA: American Society of Mechanical Engineers.

Tucker, Conrad S., & Kim, Harrison M. 2011. Trend mining for predictive product design. *Journal of Mechanical Design*, **133**(11), 111008. 00011.

Tversky, Amos, & Gati, Itamar. 1978. Studies of similarity. *Cognition and categorization*, **1**(1978), 79–98. 00504.

Tversky, Amos, & Hutchinson, J. 1986. Nearest neighbor analysis of psychological spaces. *Psychological Review*, **93**(1), 3. 00167.

University of Michigan - Optimal Design Laboratory. 2013. *Turker design - crowdsourced design evaluation*. http://www.turkerdesign.com.

van der Maaten, Laurens. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Van Horn, David, Olewnik, Andrew, & Lewis, Kemper. 2012. Design Analytics: Capturing, Understanding, and Meeting Customer Needs Using Big Data. *Pages 863–875 of: Proceedings of the ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.

Vandenberg, Steven G., & Kuse, Allan R. 1978. Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, **47**(2), 599–604. 01509.

Von Ahn, Luis, Maurer, Benjamin, McMillen, Colin, Abraham, David, & Blum, Manuel. 2008. recaptcha: Human-based character recognition via web security measures. *Science*, **321**(5895), 1465–1468.

Von Neumann, John, & Morgenstern, Oskar. 2007. *Theory of games and economic behavior*. Princeton university press. 26014.

Waggoner, Bo, & Chen, Yiling. 2013. Information elicitation sans verification. *In: Proceedings of the 3rd Workshop on Social Computing and User Generated Content*.

Wainwright, Martin J, & Jordan, Michael I. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, **1**(1-2), 1–305.

Warnaar, Dirk B, Merkle, Edgar C, Steyvers, Mark, Wallsten, Thomas S, Stone, Eric R, Budescu, David V, Yates, J Frank, Sieck, Winston R, Arkes, Hal R, Argenta, Chris F, *et al.* 2012. The Aggregative Contingent Estimation System: Selecting, Rewarding, and Training Experts in a Wisdom of Crowds Approach to Forecasting. *In: Proceedings of the 2012 AAAI Spring Symposium: Wisdom of the Crowd.*

Wassenaar, Henk Jan, & Chen, Wei. 2001. An approach to decision-based design. *ASME Paper No. DETC01/DTM-21683.* 00060.

Wauthier, Fabian L, & Jordan, Michael I. 2011. Bayesian Bias Mitigation for Crowdsourcing. *Advances in Neural Information Processing Systems*, 1800–1808.

Welinder, Peter, Branson, Steve, Perona, Pietro, & Belongie, Serge J. 2010a. The multidimensional wisdom of crowds. 00197.

Welinder, Peter, Branson, Steve, Belongie, Serge, & Perona, Pietro. 2010b. The Multidimensional Wisdom of Crowds. *Advances in Neural Information Processing Systems*, **10**, 2424–2432.

Whitehill, Jacob, Ruvolo, Paul, Wu, Tingfan, Bergsma, Jacob, & Movellan, Javier. 2009a. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems*, **22**(2035-2043), 7–13. 00234.

Whitehill, Jacob, Ruvolo, Paul, Wu, Tingfan, Bergsma, Jacob, & Movellan, Javier R. 2009b. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Advances in Neural Information Processing Systems*, **22**, 2035–2043.

Wu, Zhichao, & Duffy, Alex HB. 2004. Modeling collective learning in design. *AI EDAM: Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **18**(04), 289–313. 00016.

Yan, Xinchen, Yang, Jimei, Sohn, Kihyuk, & Lee, Honglak. 2015. Attribute2Image: Conditional Image Generation from Visual Attributes. *arXiv preprint arXiv:1512.00570.*

Yang, Maria C. 2010. Consensus and single leader decision-making in teams using structured design methods. *Design Studies*, **31**(4), 345–362.

Yannou, Bernard, Dihlmann, Markus, & Awedikian, Roy. 2008a. Evolutive Design of Car Silhouettes. *Pages 15–24 of: ASME 2008 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.* New York, NY: American Society of Mechanical Engineers. 00010.

Yannou, Bernard, Dihlmann, Markus, & Awedikian, Roy. 2008b. Evolutive Design of Car Silhouettes. *Pages 15–24 of: ASME 2008 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.* New York, NY: American Society of Mechanical Engineers.

Yin Wong, Ho, & Merrilees, Bill. 2008. The performance benefits of being brand-orientated. *Journal of Product & Brand Management*, **17**(6), 372–383. 00075.

Yumer, Mehmet Ersin, Asente, Paul, Mech, Radomir, & Kara, Levent Burak. 2015a. Procedural Modeling Using Autoencoder Networks. *Pages 109–118 of: Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM.

Yumer, Mehmet Ersin, Chaudhuri, Siddhartha, Hodgins, Jessica K., & Kara, Levent Burak. 2015b. Semantic shape editing using deformation handles. *ACM Transactions on Graphics (TOG)*, **34**(4), 86.

Yumer, Mehmet Ersin, Chaudhuri, Siddhartha, Hodgins, Jessica K., & Kara, Levent Burak. 2015c. Semantic shape editing using deformation handles. *ACM Transactions on Graphics (TOG)*, **34**(4), 86. 00003.

Zaidan, Omar F, & Callison-Burch, Chris. 2011. Crowdsourcing translation: Professional quality from non-professionals. *Pages 1220–1229 of: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Zeiler, Matthew D., & Fergus, Rob. 2014. Visualizing and understanding convolutional networks. *Pages 818–833 of: Computer Vision ECCV 2014*. Springer.

Zeithaml, Valarie A. 1988. Consumer Perceptions of Price, Quality, and Value: A Means-End Model and Synthesis of Evidence. *Journal of Marketing*, **52**(3), 2–22. 10058.

Zhang, Binbin, & Rai, Rahul. 2014. Materials Follow Form and Function: Probabilistic Factor Graph Approach for Automatic Material Assignments to 3D Objects. *Pages V007T07A012–V007T07A012 of: ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers.