# Modeling expert choice at the technical frontier

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Engineering & Public Policy

**Patrick Funk**

B.A., Mathematics, University of Montana
M.S., Engineering and Public Policy, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA
December, 2020

# Acknowledgements

First and foremost, thank you to my committee: Alex Davis (chair), Parth Vaishnav, Aarti Singh, Baruch Fischhoff, and Barry Dewitt. Alex's passion for uncovering and testing new ideas is contagious. I have walked into his office many times in the past years feeling stuck and defeated (as is often the state of a PhD student), and he is continually able to remind me of the wild joy within the process of discovery. In more ways than I can express here, Alex deserves all the credit for my completion of this thesis. I am forever grateful. Thanks to Parth who has also been a part of this journey from the start. Parth's mix of unflappable demeanor and willingness to share his expertise is inspiring and part of the most fulfilling times in my early work. Thank you to Aarti for allowing me to repeatedly run into her office, grab a student or two, pitch an idea, only to have it fall apart due to the many limitations that arise in research plans. Though those projects did not "produce" as hoped, the intellectual value I gained collaborating across the theoretical and practical elements of different disciplines was immense. As the 11th hour addition to the committee, thank you to Baruch for your willingness to dive in on short notice. More importantly, Baruch opened his home to me and my fellow students in times of celebration and deep mourning. His contribution to the community within EPP holds a special place in my experience. Lastly, Barry has been a mentor through this process since my visit day, who I am glad to say has also become a friend. Barry is the type of colleague every student wants, but few are lucky enough to have, to shadow and emulate. Barry has had innumerable contributions to this process, but his expertise in LaTeX is particularly appreciated here as his custom package is used in the formatting of this document.

Thank you to my funding sources in my time at Carnegie Mellon: The Manufacturing Futures Initiative and the NASA University Leadership Initiative. Without generous support from

organizations like these I would not have been able to attend Carnegie Mellon University. I am indebted and welcome the opportunity to pay it forward.

Thank you to my family and friends who have supported me through this process. Though their contributions are too numerous to list, my community is fundamental to any success I have had in this or any other process. They are a well-spring of love and support, allowing me to take risks and to fail spectacularly as I grow intellectually and personally. My mother (Jeanne), my father (Peter), and my brother (Andrew) don't deserve to be subjected to holiday vacation practice presentations or panicked calls about deadlines or qualifiers or any of the many things I have called them in a panic about. But, I am lucky and grateful that despite their not deserving it, they have put up with it all with a smile and at least feigned interest. I am endlessly fortunate to have you as my family.

Finally, to the many early educators who set me on this path, thank you. To Karl Wolf, Donna Turcott, and Tina Veroulis, thank you for instilling a life-long joy for learning. To John George, thank you for challenging me to embrace the discomfort of not knowing. And thank you to everyone at Grandstreet Theatre. The lessons I learned there continue to be foundational to all my proceeding endeavors.

# Abstract

Commercialization of a new material or process invention can take decades. A predominance of tacit knowledge, information asymmetries, and insufficient human capital with knowledge in the field can contribute to this delay. At this *technical frontier*, expert decision making drives critical research and development. By surveying the top experts in an industry that exemplifies the technical frontier, metal additive manufacturing for aerospace, I find that there is inconsistency in decision making structure and process both within and across experts. This inconsistency leads to an investigation of how to evaluate accuracy in a field when gathering outcome data is expensive or only exists in the future (as is the case at the technical frontier). I prove bounds on expert accuracy that rely on response structure without requiring outcome data for validation. The inconsistency in expert decision process seen in the survey case study motivates the search for a single tool to learn a diverse set of behavioral decision making models. I create a novel neural network that is able to learn some, but not all, of a variety of common theoretical behavioral choice models. This combination of qualitative and quantitative modeling of decision making at the technical frontier provides tools for speeding the development of new technologies across the "valley of death" on their way to mass commercialization.

# Contents

# List of Tables

# List of Figures

# 1

# Introduction

The technical frontier is technological development stage between the basic research that a technology is founded on and the mass adoption of that technology (Bohn, 2005). This development stage is characterized by a small number of experts who rely largely on knowledge that is difficult to codify or teach (Crane, 1969; Price, 1963; Hayek, 1945; Ackerloff, 1970; McNichols, 2008). This *tacit* knowledge is often transferred slowly through apprenticeship-style training in labs and research teams and may take decades to diffuse out into the economy at large (Evers, Cunningham, & Hoholm, 1988). The pace of innovation is therefore reliant on the highly uncertain and difficult to codify research and development choices of this small expert cohort. This combination of a challenging decision making environment and the large potential impact of those decisions is what motivates my study of choice in this context.

In the following studies I examine decision making at the technical frontier from a variety of angles. First, I present a case study utilizing experts in metal additive manufacturing for aerospace making predictions about which parts are more feasible for transition from traditional manufacturing methods to this new technology within the next five years. I separate components of their tacit and explicit knowledge and model their preferences under these individual and combined knowledge modes. Second, I present a theoretical proof on bounds on the accuracy of a respondent in a pairwise comparison task that can be applied *without* information about the ground truth outcomes of the task. This may be particularly useful at the technical frontier, where outcome information is often expensive to obtain or only exists in the future. Lastly, I examine a

novel machine learning architecture for it's ability to learn a diversity of behavioral choice models without relying on the typical theoretical assumptions of those models. Combined, this thesis presents a framework and set of tools for measuring and modeling choices at the technical frontier.

# 2

# Individual inconsistency and aggregate rationality: Overcoming inconsistencies in expert judgment at the technical frontier

## 2.1 Introduction

Going from material or process invention to commercialization often takes more than 20 years (Council, 2004; An & Ahn, 2016). Contributing to this "valley of death" (Rogers, 1962) are information asymmetries (Ackerloff, 1970; Hayek, 1945), lack of experts with knowledge in the new field (Crane, 1969), the predominance of tacit knowledge (Bonnín Roca, Vaishnav, Morgan, Mendonça, & Fuchs, 2017; Polanyi, 1966; Price, 1963), and few investors willing to fund risky new approaches (Russell & Fielding, 2014). Particularly challenging in these contexts can be transitioning from "art" to "science" (Bohn, 2005). In this work we leverage an important emerging technology, metal additive manufacturing, that offers an extreme example of such issues. Our approach aims to characterize expert decision-making at the technological frontier, then explore opportunities for interventions that accelerate new technology commercialization.

Although experts are critical for the diffusion of new technologies, errors in expert judgment can prevent the effective use and transfer of expert knowledge. Research shows that human decision-making can be inconsistent – both across and within people (Tversky, 1969). This inconsistency can come from a number of possible sources: the type of problem, the judge's decision rule, uncertainty in the judge's knowledge, or randomness in judgments (Thurstone, 1927). For example, process of elimination decision rules can lead to inconsistent rankings of alternative courses of action (Tversky, 1972). In general there are three types of decision rules employed by experts: tree-based (R. D. Luce, 1956; Tversky, 1972; Tversky & Sattath, 1979; Batley & Daly, 2006), weighted linear additive rules (McFadden et al., 1973), and non-linear rules. Understanding which experts employ each decision rule is critical to evaluating the quality of expert advice derived from the rule. Expert decision-making will also exhibit random variation when alternatives are similar (Marschak, 1959), and systematic variation depending on how quickly experts must make choices (Busemeyer & Townsend, 1993) or when decision processes are context-dependent (Tversky & Simonson, 1993; Tversky & Kahneman, 1986, 1981). In small samples – for an emerging technology where there are few experts – small amounts of inconsistency can lead to large deviations from optimal commercialization paths.

In what contexts expert insights outperform statistical models, and vice versa, remains an open and important direction of inquiry, as well as a moving target. In the 1970's and 1980's, Dawes found linear statistical models performed better than experts at prediction across many contexts (Dawes & Corrigan, 1974; Dawes, 1979). In analyzing his findings, Dawes suggests that experts can accurately choose decision criteria, but can be inferior to statistical models when it comes to aggregating that information into a prediction (Dawes, Faust, & Meehl, 1989). While statistical decision making has outperformed experts in many domains (Sawyer, 1966; Grove, Zald, Lebow, Snitz, & Nelson, 2000), recent research suggests that a combination of expert and statistical insights can lead to better prediction accuracy than when either is used in isolation (WA, FE, J, & et al, 1995; Torrano-Gimenez, Nguyen, Alvarez, & Franke, 2015; Abdollahi, Davis, Miller, & Feinberg, 2018; Holzinger, 2016). Expert judgments themselves can also be improved with carefully designed interventions and training (Kadane & Fischhoff, 2013; Morgan, 2014; Mellers et al., 2014).

Much of the literature on expert decision-making has focused on the application of existing knowledge in professional contexts, such as medicine, clinical psychology, or criminology (Sawyer, 1966; Dawes et al., 1989; Grove et al., 2000; Garb, 1989; Sinuff et al., 2006). Less is known about expert decision-making at the frontier of new technologies. In the context of scientific discovery, Dunbar (Dunbar, 1997) finds that scientists doing research at the boundaries of human knowledge simplify problems using heuristics and analogies, similar to the ways laypeople simplify problems in other contexts (Tversky & Kahneman, 1974). In the context of applying existing knowledge to new problems, McComb finds that engineering students use a series of general design processes, and that optimizing the order of these design steps can yield improved performance results for the final product (McComb, Cagan, & Kotovsky, 2017). Qualitative expert judgment is used in technology forecasting, but often as a guiding step in bibliolgical work (Robinson, Lagnau, & Boon, 2018).

In this work we characterize the judgment of experts at the technological frontier, specifically engineers judging the suitability of components for production using metal additive manufacturing. This work is an example of use-inspired basic research, addressing fundamental questions of how decision-making takes place in the context of uncertainty inherent to the technical frontier, with the goal of speeding the transition through this bottleneck for future technologies (Stokes, 2011). In our context, the decisions are as much art as science, where the underlying decision processes used to determine which parts should be made using additive manufacturing has yet to be standardized, and cannot currently be characterized by formal scientific models. Additionally, the design that would be used to produce the component with the new technology and which components are best made with that new technology remains a topic of debate, where the correct answer is unknown.

Due to these challenges at the technical frontier, there are few successful application cases on which to base expert accuracy. Traditionally, combining individuals' knowledge or preference falls into two categories, those with the goal of maximizing combined social welfare, and those attempting to predict most accurately on an unknown task. At the technical frontier, we are not balancing preferences, but rather trying to understand the expertise required to move the technology forward more rapidly. Theorems of social utility, such as Arrow's Impossibility Theorem, are not about aggregation of knowledge, but instead about balancing preferences.[1] Past

[1] Arrow's Impossibility Theorem relies on two key factors, which are not relevant to our context. First, we do not argue that the aggregate preference order needs to be transitive as is required in Arrow's case, but rather that its form (even the intransitive form in our data) is valuable over the variability in order and structure of the individual experts. In the extreme case (with further verification of preferences mapped to technology developments) it may even be desirable for the expert group to have a dictator (weighting the expert with the most accurate knowledge (?, ?)), which is in direct conflict with Arrow's axioms.

work on wisdom of the crowds and expert forecasting have examined the cases where accuracy is known and constant, but do not address knowledge at the technical frontier where there is significant uncertainty, what is accurate may be unknown, and the state of knowledge potentially time-varying. This distinction is important: While there are a number of similarities between preferences and judgements because they share common axioms that apply to individuals, such as transitivity, criteria that apply to aggregate utility or aggregate preferences need not be related to the criteria that apply to aggregate judgements about technology feasibility.

We focus on the extreme example of metal additive manufacturing (MAM) in aerospace. MAM in aerospace serves as a useful case to study in that many of the factors that create challenges in the commercialization of a new technology are represented in particularly challenging forms. Tacit knowledge is known to be particularly important in the aerospace industry (McNichols, 2008). This high tacit knowledge content may make for particularly slow diffusion of knowledge within or between firms. In this already challenging context, the number of experts with the scientific or technological knowledge at the technical frontier of MAM are not necessarily the same as those with MAM-relevant design knowledge nor are these the same as those experts with hands-on MAM production experience and capabilities. The experts with aerospace-relevant knowledge, such as in aircraft body or engine design are again a different set. The experts with overlapping knowledge across even two of these categories can likely be counted on one hand. Competitive pressures have also limited knowledge flows between regulators, large OEMs, and smaller suppliers of repair parts in the industry (Bonnín Roca et al., 2017). Finally, the possible negative effects to civilians and associated set-backs to technological progress of an MAM-driven failure in the context of aerospace applications are likewise extreme: loss of human life due to a technical problem in a civilian aircraft can lead to long term setbacks in technology adoption and diffusion as shown in the case of the de Havilland Comet jet aircraft window failures (Roca, Vaishnav, Fuchs, & Morgan, 2016).

In addition to its usefulness as a case study, MAM is important both economically and for national security. MAM's most promising early applications are airframes and jet engines, civilian exports of which constitute the largest category of US exports (Bureau, 2017). Other promising applications range from the automotive industry to oil and gas to a broad range of military products. From a national security perspective, it is impossible to quantify the cost of not having a technologically superior weapons system in a time of military crisis. Given the above applications, the federal government has been making significant investments in accelerating the commercialization MAM: In 2014, the US federal government invested $40 million in additive manufacturing through America Makes (formerly the National Additive Manufacturing Innovation Institute) with an additional $50M in matching state and local government, philanthropic, and corporate funding (Russell & Fielding, 2014). As with the other national manufacturing institutes, these funds are focused on technology readiness levels 4-7, in theory to help bridge the valley of death.

## 2.2 Methods

### 2.2.1 Research design

Our research design uses a two-stage approach. In our first stage we conducted informal interviews to elicit the part attributes that experts believe would most strongly determine a part's feasibility for metal additive manufacturing in the next five years. To obtain different perspectives without exhausting our population of experts, we interviewed one expert from each of academia, government, and industry. From these interviews we identified a set of eight attributes associated with part feasibility for MAM that spanned the space of attributes expressed by these three experts: criticality, tolerances, overhangs, non-cyclic thermal load, cyclic thermal load, non-cyclic mechanical load, cyclic mechanical load, and the part's economic feasibility. We then chose parts that ranged from those currently produced with MAM and certified by the FAA (an engine bracket), to parts which our preliminary expert group believed were impossible to manufacture by MAM in the near future (a turbofan gearbox). Based on these criteria and suggestions from the three experts, we selected twelve civil aircraft engine parts that varied in their feasibility for MAM: an engine bracket, a compressor blisk, a fan disk, a thrust reverse blocker door, a monocrystalline turbine blade, a turbofan gearbox, a heat exchanger, an inlet cone, a fuel swirler, vanes, an annular combustor, and combustion chamber lining. We chose parts for the survey solely from jet engines, rather than including airframe components, in order to narrow the scope of the part choices. We found images of each part and constructed definitions of each of the eight part attributes, then pilot tested those parts and attribute questions for clarity and construct validity with the original three experts and colleagues knowledgeable about the topic (Shadish, Cook, & Campbell, 2002). Descriptions of the parts can be found in SI Section 1.

In the second stage we surveyed a broader sample of experts to estimate our quantitative models. The survey consisted of two sections: a pairwise comparison task and an explicit rating task. By capturing these two measures of the same part set we aimed to isolate not only tacit and explicit knowledge, but also the internal consistency and decision process of experts in order to uncover the implications of information asymmetries and small expert groups. The pairwise comparisons consisted of repeated two alternative forced-choices[2] that asked experts to choose between pairs of parts based on their feasibility for MAM, allowing us to understand how experts make decisions based on their knowledge of the parts, rather than our assumptions about their knowledge. Experts were shown pairs of the 12 engine parts along with their physical dimensions. The order of part pairs was randomized across study participants. For each pair of parts, experts were asked ”Which part has the greater likelihood of being feasible for metal additive manufacturing within the next 5 years?” *Feasibility* was explicitly defined in the instructions as the ability for a

> part to be produced with equivalent or better functionality and ready for certification using metal additive manufacturing within the next 5 years. We are interested here both in part performance, as well as if there is an economic case for its production. Please consider the potential for part redesign. You do not need to be limited to existing off-the-shelf equipment or materials.

The experts responded by selecting one of the two parts from each pair, as shown in Figure 6.1, top panel. Our feasibility definition focuses on whether the part could achieve mass adoption using MAM within five years, taking into account the pace of change in the industry, the importance of both technical details and economic value for development pipelines, the likely benefits of redesign enabled by MAM, and the rapid growth in machine and raw material options.

[2] While there are examples of indifference options being allowed in designs such as ours (Regenwetter & Davis-Stober, 2008), we excluded this option and forced respondents to make a choice in each comparison because indifference is rarely chosen and can alternatively be expressed as choice probabilities of 0.5 (in repeated choices) (Marschak, 1959).

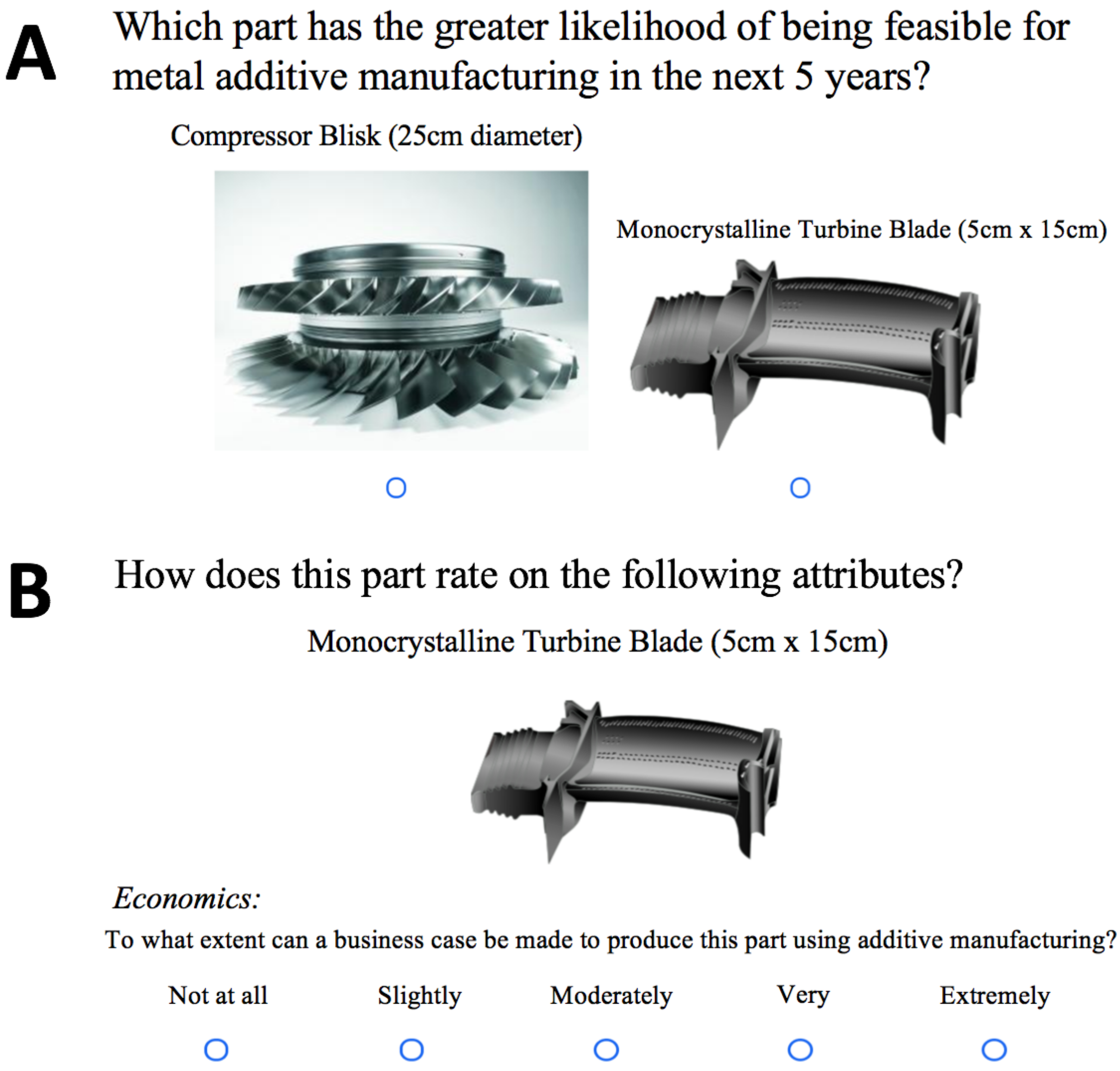
The explicit rating task, as shown in the bottom panel of Figure 6.1, asked experts to score each part on a 1 to 5 scale based on the eight attributes that were most important in the informal interviews that we conducted in the first stage of our research (i.e., criticality, economics, etc.). The pairwise comparison task always came before the explicit rating task because we expected an asymmetrical transfer effect, where one task affects the performance of a subsequent task but not vice versa (C Poulton & R Freeman, 1966). In this case, we expected that putting the explicit rating task first would unnaturally focus participants on those eight attributes in the pairwise comparison task, but that there would be no transfer effect when the pairwise comparison task came first.

Lastly, respondents were asked to report their age, the number of years they have worked in academia, government, and industry, their expertise in MAM and aerospace, and their confidence in their explicit judgments.

To ensure that respondents understood the task and were paying attention, the pairwise comparison had two checks. Each check paired a part that is known to be feasible, General Electric's (GE) fuel nozzle, against what the preliminary experts agreed was a part impossible to make with additive manufacturing in the foreseeable future, a turbo fan gearbox. These two parts were paired against each other as the first and last question of the pairwise comparison task. Failure of either of these questions (choosing the turbofan gearbox over the fuel nozzle) was used as removal criterion from the planned analysis, but these experts were included in subsequent sensitivity analyses. (The preregistered analysis is available on the Open Science Framework – `osf.io/djbcm`).

### 2.2.2 Data collection and sample

Our survey population is experts in aerospace, metal additive manufacturing, or a combination of these fields. We aimed to contact all experts in MAM for jet engine applications. We began with a small group of expert contacts from the researchers' networks. We expanded the target group by searching through the editorial boards of professional and academic journals, widely published and cited academic papers, leading positions at prominent companies in the field, and government or military staff within additive and aerospace functions. To grow the list of experts further, we used a snowball method asking respondents to recommend two or three peers who would be willing to participate. Overall we assembled a target group of 67 experts. We contacted members of



Which part has the greater likelihood of being feasible for metal additive manufacturing in the next 5 years?

Compressor Blisk (25cm diameter)

Monocrystalline Turbine Blade (5cm x 15cm)

○ ○

How does this part rate on the following attributes?

Monocrystalline Turbine Blade (5cm x 15cm)

*Economics:*

To what extent can a business case be made to produce this part using additive manufacturing?

| Not at all | Slightly | Moderately | Very | Extremely |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

Figure 2.1: Top: Example of the pairwise comparison task in the first section of the survey (GE Aviation). Bottom: Example of the explicit rating task in the second section of the survey (Wikimedia Foundation). All eight questions can be found in SI Section 2.

the target expert group over email and phone to request their participation. Initially, we emailed experts monthly to invite their participation. In the last 3 months of data collection we emailed and called experts weekly. Once an expert accepted the invitation to join the study, we sent a link to an online survey. For participants who were willing to call in while completing the online survey, a researcher listened and recorded comments from the participants, but provided no clarification or additional information. For 20 experts who volunteered, a researcher conducted a brief interview to discuss the survey results and provide qualitative context for the responses after the survey was completed. We collected surveys from a total of 39 experts out of our 67 targeted experts for a response rate of 58%. However, 27 experts completed the survey and passed all attention and expertise checks, giving a completion rate of 40%. Of the 12 surveys which were not included in the analysis, nine were incomplete and three failed the attention/expertise checks within the survey. Experts' self-identified experience covers industry, academia and government as well as every combination of these three.

Table 2.1: Main Sample Demographics

| | |
|---|---|
| Contacted | 67 |
| Only First Half Completed | 3 |
| Passed Attention/Expertise Checks | 2 |
| First and Second Half Completed | 29 |
| Passed Attention/Expertise Checks* | 27 |
| *Expertise Represented in Main Sample* | |
| Industry | 3 |
| Academia | 2 |
| Government | 1 |
| Industry + Academia | 9 |
| Industry + Government | 1 |
| Academia + Government | 1 |
| Industry + Academia + Goverment | 8 |

*Survey responses used for main sample

### 2.2.3 Methods for analysis

The two-stage survey design allows us to examine expert judgment from several complementary viewpoints. Pairwise comparisons allow us to determine each expert's part ranking in terms of its feasibility for additive manufacturing, as well as whether that ranking is consistent (transitive). If an expert's judgments are transitive, then the highest ranked part should be chosen in all 11 paired comparisons in which it appears, the second highest ranked should be chosen in 10 pairs – all save the one where it is paired against the highest ranked part – the third highest ranked should be chosen in 9 pairs, *etc.* Inconsistencies (i.e., intransitivities) arise whenever part A is selected over B, B over C, and C over A. In this case, there is ambiguity in how to represent the individual's choices, as they cannot be represented by a ranking. To handle this we use a tallying method, where parts are ordered by the total number of times each part was chosen in its comparisons (Goddard, 1983).

Pairwise comparisons also allow us to evaluate the choice process of experts in aggregate using multidimensional scaling (MDS) (Torgerson, 1958). MDS is a dimension reduction method that takes dissimilarity data as an input and returns a set of points in a lower-dimensional space that most closely approximates the original dissimilarities. In the pairwise comparison task, two parts are most dissimilar if, across all the experts, one is always chosen over the other (with choice probabilities $p_{ij}$'s of zero or one), and most similar if one part is chosen over the other half the time. Thus, the dissimilarity transformation is defined by $\delta_{ij} = |p_{ij} - 0.5|$, where $p_{ij}$ is the proportion of experts that chose part $i$ over part $j$ as being more feasible for additive manufacturing. The $\delta_{ij}$ index ranges from 0 (most similar, $p_{ij} = 0.5$) to 0.5 (most dissimilar, $p_{ij} = 1$ or 0). Given a matrix of these dissimilarities, MDS finds a low-dimensional approximation of the matrix that minimizes the *stress-1* metric (Kruskal, 1964):

$$\sigma_1 = \sqrt{\frac{\sum[f(\delta_{ij}) - d_{ij}(\mathbf{X})]^2}{\sum d_{ij}^2(\mathbf{X})}}. \tag{2.1}$$

Here, $d_{ij}$ is the distance between parts $i$ and $j$ in the low-dimensional representation $\mathbf{X}$. The mapping $f$ on the dissimilarities $\delta_{ij}$ depends on the scale-type of the data (e.g., interval, ordinal). An iterative procedure is used to search for configurations $\mathbf{X}$ to minimize $\sigma_1$. One can also lower stress by choosing different models for $f$, although that requires making different (and possibly stronger) assumptions about the dissimilarity data (Borg & Groenen, 2005).

Where $\delta_{ij}$ is the dissimilarity between parts $i$ and $j$, in our case the difference between the average number of times part $i$ was chosen over part $j$, $d_{ij}$ is the distance in the new, reduced-dimensional space between parts $i$ and $j$, and $\mathbf{X}$ is the current $m$-dimensional configuration of the MDS fit. We use a majorization algorithm for the MDS estimation, implemented in the SMACOF package in R (de Leeuw & Mair, 2009).The advantage of MDS is that if a one-, two-, or three-dimensional MDS solution fits the data well, then the data can be visualized.

We compare the MDS dimensions to the explicit attribute ratings from the survey. Because MDS results are invariant under rotations, translations, and expansions, we use an algorithm that finds a best fit between the MDS dimensions and explicit attribute ratings. This process is known as a Procrustean analysis (Hurley & Cattell, 1962), where the congruence coefficient (Tucker, 1951) $c = \frac{\sum_i x_i y_i}{\left[\sum_i (x_i^2) \sum_i (y_i^2)\right]^{\frac{1}{2}}}$ (also known as the cosine similarity), between the explicit and MDS dimensions is minimized. The variables $x_i$ and $y_i$ are distances between parts in the MDS transformed space and in the explicitly rated part space.

A third method, discrete choice modeling (DCM), blends the explicit and pairwise comparison parts of the survey. In the discrete choice model we estimate choice probabilities for each pair of parts as a function of the explicit part attribute ratings using the binary logit model (see (2.3)) (McFadden et al., 1973). Additionally, we model taste variation among the expert set using a mixed logit model (Train, 2009). While MDS uses average response data to do dimensional reduction, a DCM uses individual responses to learn a weighted additive function that best explains respondents' choices. For a discussion of examples where the stochastic additive model assumptions are not met (e.g., single attribute decision criteria, decision tree structures, nonlinear decision models) see SI Section 3. Based on our initial discussions with experts, we hypothesized a model that included criticality, tolerances, overhangs, and economics (see link above for preregistered analysis). In this model, the propensity $V_{it}$ of participant $i$ to choose the alternative shown on the left (L) over the right (R) on trial $t$ is the weighted sum of the difference in the explicit ratings the participant gave to the two alternatives, where the unknown weight vector $\beta$ is estimated using maximum likelihood estimation:

$$\begin{aligned} V_{it} =& \beta_1 \Delta Criticality_{it} + \beta_2 \Delta Overhang_{it} + \\ & \beta_3 \Delta Tolerance_{it} + \beta_4 \Delta Economics_{it} \end{aligned} \tag{2.2}$$

For example, the difference in the explicit ratings of criticality between the left and right parts shown on trial $t$ to participant $i$ is denoted $\Delta Crit_{it} = Crit_{itL} - Crit_{itR}$. Choice probabilities are related to $V_{it}$ through the logistic function:

$$P_{it} = \frac{e^{V_{it}}}{1 + e^{V_{it}}} \tag{2.3}$$

where $P_{it}$ is the probability of choosing the part on the left for participant $i$ on trial $t$. Our second discrete choice modeling approach is the mixed logit model, which allows for participant-level heterogeneity in the weight placed on each attribute difference:

$$\begin{aligned} V_{it} =& \beta_{1i} \Delta Criticality_{it} + \beta_{2i} \Delta Overhang_{it} + \\ & \beta_{3i} \Delta Tolerance_{it} + \beta_{4i} \Delta Economics_{it} \end{aligned} \tag{2.4}$$

where $\beta_i = [\beta_{1i}, \beta_{2i}, \beta_{3i}, \beta_{4i}]$ are random variables with a multivariate normal distribution with mean vector $\mu = [\mu_1, \mu_2, \mu_3, \mu_4]$ and variance-covariance matrix $\Sigma$. Choice probabilities are related to $V_{it}$ through the logistic function:

$$P_{it} = \int_{\beta} \frac{e^{V_{it}}}{1 + e^{V_{it}}} f(\beta) d\beta \tag{2.5}$$

where $\int_{\beta} \dots f(\beta) d\beta$ is a multiple integral over the multivariate distribution of the $\beta$ vector. Estimation was done using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015).

## 2.3 Results

### 2.3.1 Part rankings

First we examine the rankings implied by experts' choices in the pairwise comparison task. As shown in Figure 2.3, after summing the adjacency matrices across all experts, we find that the aggregate ordering of parts was almost transitive, with only a single *cycle* (where experts prefer part A to B, B to C, but then C to A). Validating the approach, the parts or sub-components that are currently approved by the FAA and used in commercial aviation (the engine bracket and fuel swirler, which is a component of a fuel nozzle) were ranked at the top, while some of the most difficult parts to produce using traditional methods (turbofan gearbox and monocrystalline turbine blade) were ranked at the bottom. The single cycle occurred in the middle of the aggregate ranking, where the heat exchanger was preferred to the annular combustor (on aggregate), the annular combustor was preferred to the inlet cone, but the inlet cone was preferred to the heat exchanger.

In contrast, Figure 2.3 shows that individual experts had significant inconsistencies in their expressed judgments, where only 7/27 had no cycles. One measure of the extent of inconsistency in ordering the parts is the size of the minimum feedback edge set (MFES), which is the minimum number of edges in a graph that need to be removed to make it acyclic (Kamae, 1967). In our case, the MFES tells us how far away a given expert is from being internally consistent (having no cycles in their judgments). The average MFES is 1.7 while the median is 2. The minimum is 0 (seven experts exhibit consistent judgments) and the maximum is 7 (the far right graph in Figure 2.3). The majority of responses are near consistent or consistent, with minimum feedback edge sets of less than 4.

We also calculate how far any particular expert's judgment graph is from the aggregate using the Hamming distance (HD), which is equal to the number of entries that must be changed in the upper triangle of a respondent's adjacency matrix to match the aggregate adjacency matrix (Hamming, 1950). No two experts have the same judgments pattern. The minimum HD is 3 and the maximum is 28 with a mean and median of 14.3 and 14, respectively. Correlation between HD and MFES is 0.47. Figure 2.2 shows that those experts who are highly inconsistent are also more

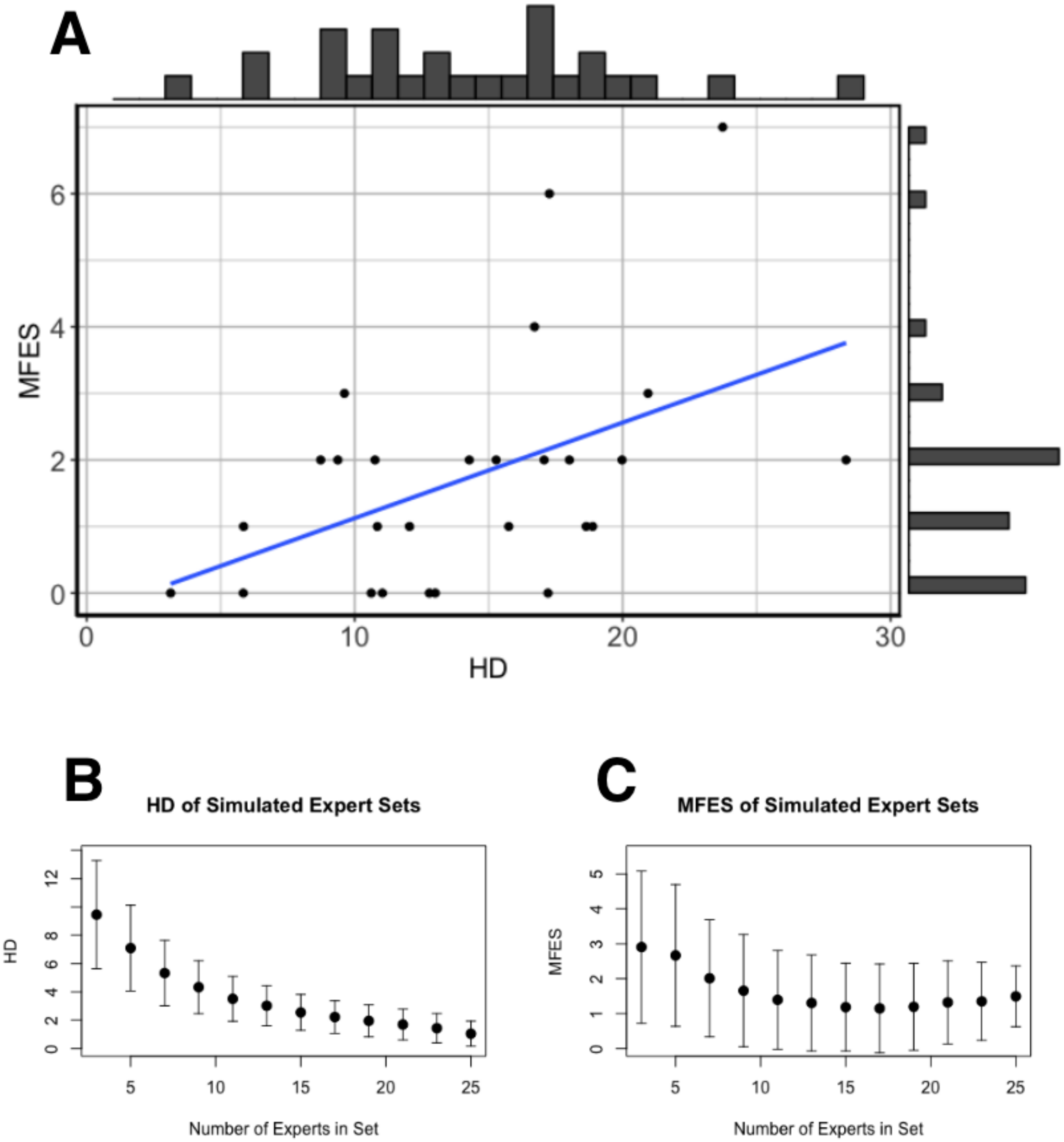


Figure 2.2: (A) Scatterplot of Hamming Distance (HD) to Minimum Feedback Edge Set (MFES) with density plots of each variable individually on the margins. A simulation study shows how (B) HD and (C) MFES vary with an increasing number of experts pooled in the aggregation process. Error bars represent one standard deviation in the mean of the 1000 samples at each group size. HD decreases steadily while MFES varies only slightly. Variance between samples decreases in both cases, however.

likely to be far away from the average judgments .

We simulated aggregation for increasing numbers of experts to see the effect of sample size on HD and MFES. Each sample was taken randomly (with replacement) from the 27 experts and the MFES and HD were calculated from the average judgments. This process was repeated 1000 times for each expert group size and the average and standard deviation of these simulations are returned in Figure 2.3. The average HD decreases as does the variance of the samples with increasing numbers of experts in the aggregation group. MFES decreases slightly, from 3.0 to 1.4, as does its variance with increasing group size. This simulation suggests a group size of 10-15 may get the greatest returns of inter- and intra- group consistency.

### 2.3.2 Multi-dimensional Scaling

Next, our multi-dimensional scaling approach found that three dimensions produced the best MDS fit to the aggregate pairwise comparison data. The stress measured at varying number of dimensions along with other validation measures of the MDS fit can be found in Appendix 1. Applying the Procrustes analysis on all combinations of three explicit dimensions returns economics, tolerances, and cyclic thermal load as the dimensions of best fit with a congruence coefficient of 0.932 (see Figure 2.4). The fuel swirler has mean explicit ratings on the dimensions of economics, tolerances, and thermal load - cyclic of (3.7, 3.6, 2.9). These values position the fuel swirler (along with the other most preferred parts) in the upper rear corner of the graph where economics are high, tolerances are low, and thermal loading is low relative to the other parts. The combustion chamber lining has mean explicit ratings on the dimensions of economics, tolerances, and thermal load - cyclic of (3.0, 3.4, 4.2) representing the middle of the order where attributes are traded off against each other. In the low feasibility range, such as the compressor blisk (2.9, 4.1, 4.5), we generally see low economic value with above average values on tolerance and thermal loading requirements. For coordinates of all parts see SI Section 6.

### 2.3.3 Discrete choice model

Fourteen of the 27 experts' choice exhibit *quasi-complete separation*, which cannot be modeled with a stochastic additive discrete choice model. For the remaining 13 experts, the hypothesized discrete

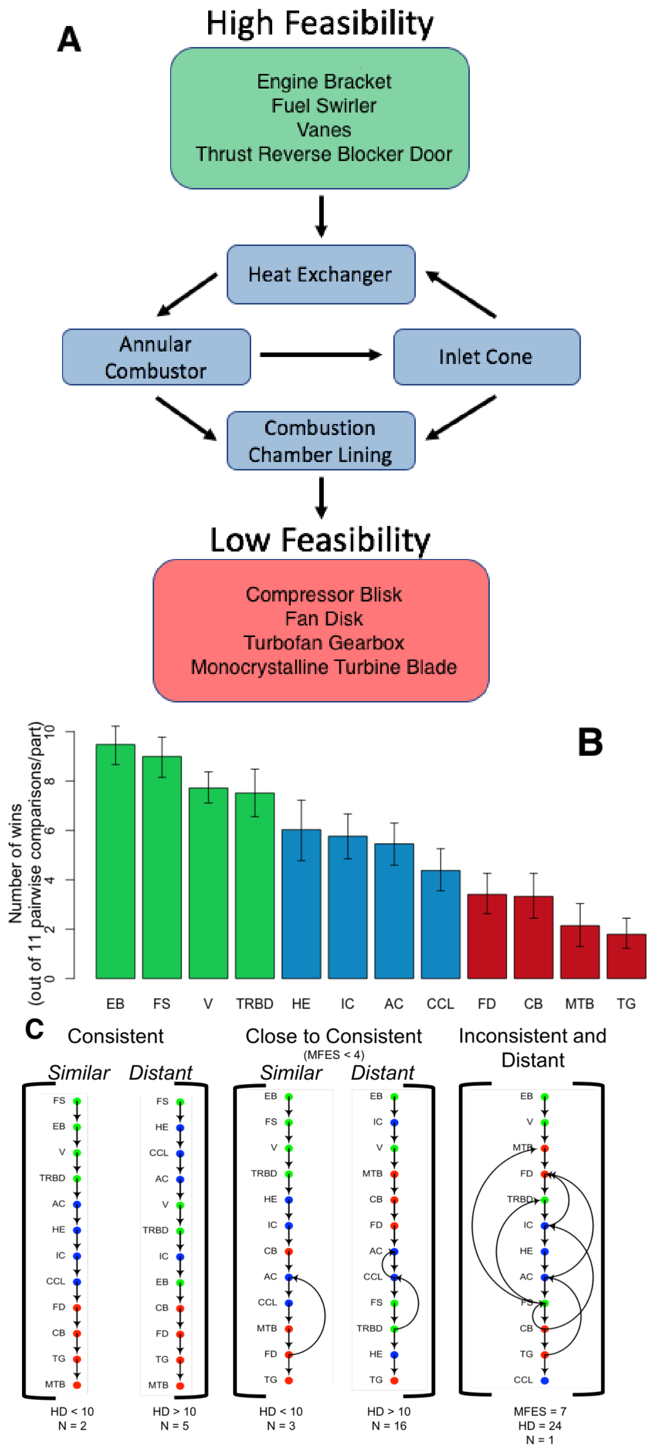


Figure 2.3: (A) Graph of the aggregate judgments across all experts for part feasibility for MAM. The graph contains a cycle within the medium feasibility region (MFES = 2). (B) Simulated expert sets are created by resampling the experts' responses with replacement in order to find bootstrapped 95% confidence intervals for the relative judgment of parts. Here this is measured by the average number of wins in a resampled pairwise comparison out of the 11 comparisons per part. EB = Engine Bracket, FS = Fuel Swirler, V = Vanes, TRBD = Thrust Reverse Blocker Door, HE = Heat Exchanger, AC = Annular Combustor, IC = Inlet Cone, CCL = Combustion Chamber Lining, CB = Compressor Blisk, FD = Fan Disk, TG = Turbofan Gear, MTB = Monocrystalline Turbine Blade. See SI Figure 5 for the statistically significant group sets (C) Five example expert judgment graphs that are representative of the different structures found in the responses. The color coding matches the categories of the above of high, medium, and low feasibility as preferred by the aggregated choice.

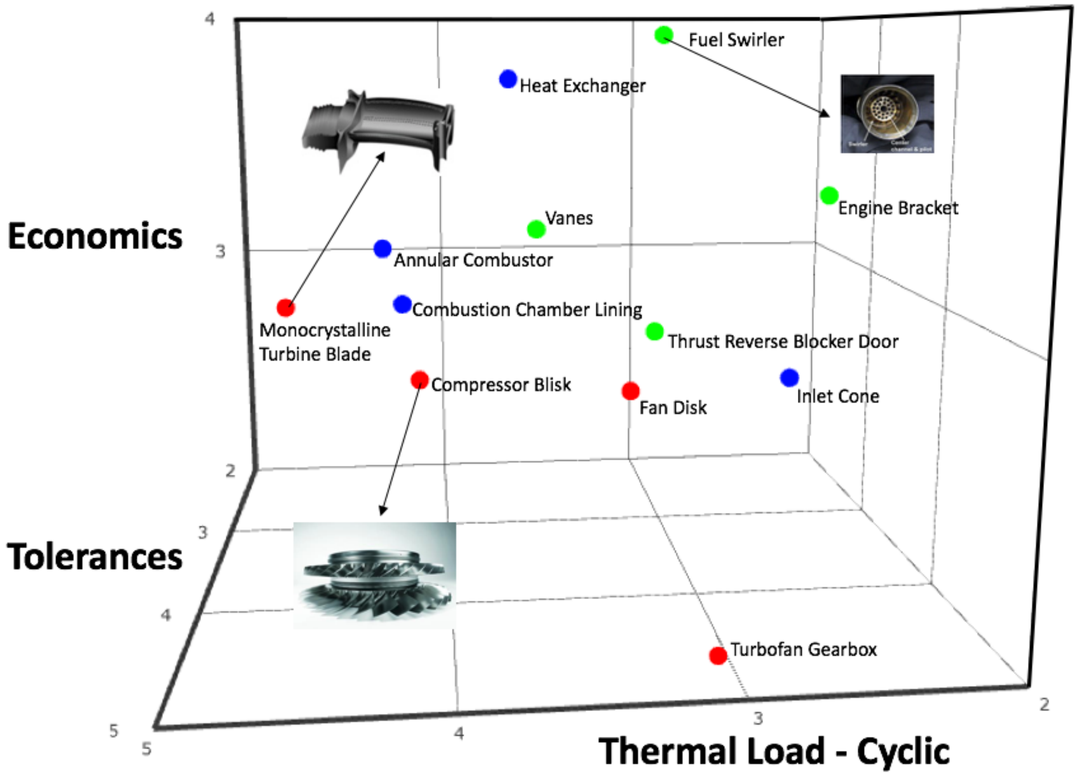


Figure 2.4: Multidimensional scaling in three dimensions modified to match the explicit ratings that fit best: economics, cyclic thermal load, and tolerances.

choice model with criticality, overhangs, tolerances, and economics as independent variables is significant in each variable (see Table 2.2). This model has a correctly classified ratio of 0.75. It should be noted again that these are the values of the differences between parts' attributes in a pairwise comparison. The signs of each variable match the hypothesized qualities. Increases in criticality, overhangs, and tolerances all reduce the feasibility of producing a part using MAM, whereas a stronger economic case increases the overall feasibility of a part for MAM. The degree of overhangs reduces the feasibility of production via MAM but its effect is lower than that of the other three variables, which are of similar magnitudes. For the model containing all variables measured, thermal load - noncyclic is an additional significant variable. Like overhangs, thermal load - noncyclic has a smaller effect than criticality, tolerances, and economics. When accounting for taste heterogeneity with the mixed logit model of the hypothesized variables, overhangs is no longer significant. Criticality, tolerances, and economics are all the same signs as the previous models and similar in magnitude. From the variance/covariance matrix, economics has nearly double the variance of all other variables. The distribution of the average part labels for these attributes are similar, suggesting that experts disagree more on economics than on criticality, overhangs and tolerances. The random effects of economics and criticality are nearly perfectly correlated, meaning as an individual weighs economics more highly, they will weight criticality more highly also, thus reducing criticality's overall weight in determining the feasibility.

Among the 14 experts who have non-additive decision rules, three exhibit single criteria decision-making. Two use only economics and one only tolerances. One expert rates all parts equally on overhangs, making it impossible to derive a $\beta_2$ from (2.2) or (2.4). The remaining 10 experts' decision rules were modeled as classification trees and an example of this process can be seen in SI Section 4. All seven experts who are internally consistent (MFES = 0) are implementing one of these simplifying decision methods.

We also ran a series of robustness checks. If the responses provided contained only noise, we would expect our MDS's stress value to be indistinguishable from a random stress value. To test this, we created randomly permuted dissimilarity matrices (consistent with experts choosing randomly between the two parts shown them) and measured the stress from a MDS of these synthesized data. The stress found from our original data lies well outside the distribution of

Table 2.2: Multinomial and Mixed Logit Choice Models

| | *Dependent variable:* | | |
|---|---|---|---|
| | y | | |
| | Preregistered<br>(Eq. 3) | All Attributes<br>(Eq. 3) | Mixed Logit<br>(Eq. 5) |
| Criticality | $-0.530^{***}$ | $-0.534^{***}$ | $-0.413^{***}$ |
| | (0.072) | (0.078) | (0.150) |
| Overhangs | $-0.130^{**}$ | $-0.122^{*}$ | $-0.127$ |
| | (0.065) | (0.067) | (0.134) |
| Tolerances | $-0.511^{***}$ | $-0.515^{***}$ | $-0.585^{***}$ |
| | (0.078) | (0.083) | (0.162) |
| Economics | $0.659^{***}$ | $0.692^{***}$ | $0.652^{***}$ |
| | (0.067) | (0.072) | (0.187) |
| Thermal Load-<br>Cyclic | | 0.128<br>(0.093) | |
| Thermal Load-<br>Noncyclic | | $-0.181^{**}$<br>(0.083) | |
| Mechanical Load-<br>Cyclic | | $-0.032$<br>(0.079) | |
| Mechanical Load-<br>Noncyclic | | 0.057<br>(0.079) | |
| Number of Decision Makers | 13 | 13 | 13 |
| Observations | 858 | 858 | 858 |
| Log Likelihood | $-435.725$ | $-433.135$ | $-428.884$ |
| Akaike Inf. Crit. | 879.450 | 882.269 | 885.768 |

Random Effects - Mixed Logit Variance/Covariance

| | Criticality | Overhangs | Tolerances | Economics |
|---|---|---|---|---|
| Criticality | 0.188 | - | - | - |
| Overhangs | 0.09 | 0.150 | - | - |
| Tolerances | 0.01 | -0.83 | 0.193 | - |
| Economics | 0.99 | 0.12 | -0.11 | 0.330 |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

stresses observed in the randomly permuted values. This allows us to reject the hypothesis that experts' choices are random. See SI Section 3 for results from permutation tests. The Procrustes analysis yields a maximum congruence coefficient of 0.932 using the explicit dimensions of economics, tolerances, and cyclic thermal load. A survey of simulations and studies suggests a congruence coefficient over 0.95 is presumed an "exact" fit in real-world data (Lorenzo-Seva & Ten Berge, 2006). The good fit to the explicit dimensions suggests that economics, tolerances, and cyclic thermal loading are important attributes in experts' choices.

### 2.3.4 Application to a new context: Proof-of-Concept

We do not have the means within this study to validate the DCM by building and testing each part. That said, we sought to apply our DCM model representing the aggregate expert's implicit and explicit judgments to a new part set to understand the ease of mapping our model variables to the type of data available in other contexts and also to sanity-check the generalizability of the model trained on one context (metal aircraft engine parts) was applied to a new one (metal aircraft parts more generally) (see Figure 2.5 for results). The aircraft part database to which we applied our model included 30 variables for each of 160 parts including size and shape parameters, materials and finish, operating conditions, and prices. In order to translate these variables to the significant DCM variables (criticality, tolerance, overhangs, and economics), we created a systematic mapping for each DCM variable.

We mapped our DCM criticality variable directly to criticality designations within the database, which had three levels. We used the International Tolerance Grade as it relates to the physical characteristics of the parts in the database to create a scaling for tolerances (ISO, 2010). We created a proxy for overhangs based on the number and complexity of joints on the part, because no visual data was provided with the database.

To construct the economic case for each aircraft part, we modified the process based cost model (PBCM) from Laureijs *et al.* 2017 (Laureijs et al., 2017) to model the part's production cost and estimate its unit production cost. We adjusted the original model (based on the GE engine bracket) to account for varying part heights (by assuming an equal time per layer and standard layer height of 20 microns for printing) as well as for the costs of additional materials including stainless steel and aluminum. While there are currently not exact alloy matches available on additive platforms for every part in the part set, stainless steel, aluminum and titanium (as the original model used) spanned a majority of the components.

We assumed that producing the part using AM will reduce weight (and therefore material use) by 35 percent (the most modest weight saving from redesigning a part for AM found by Huang *et al.* 2015 (Huang et al., 2016)). This reduction in weight factors in through a reduced material usage and therefore reduced cost in the PBCM.

We then applied the model described by (2.2), with weights derived from our expert set's explicit and tacit knowledge, to rank the aircraft parts from most appropriate for MAM to least. See Figure 2.5 for the top parts and their modeled feasibility scores. This process simplifies a complex evaluation of hundreds of part variable combinations into a single model output quantitatively derived from the aggregate knowledge of the top experts in the field. As can be seen in the figure, the highest feasibility parts according to the model were rails involved in loading equipment into the aircraft. This result passes a basic sanity test, as the potential weight savings are large (implying direct fuel cost savings and indirect benefits of extended transport through light-weighting), their geometric constraints are not limiting, and their criticality is low.

As in any attempt to use a model beyond the scope of its creation, there are limitations. In this case, we assigned values to our DCM variables from the aircraft database in a systematic but *ad hoc* manner. Ideally this process would be informed by expert input or learned through expert models. One element of this assignment involved the techno-economic modeling of these parts. This PBCM model was built for the production process of a specific part, the GE engine bracket, and needs further refinement to fully capture the diversity of parts presented in the aircraft part set. Given the unique requirements of subsets of aircraft applications (freight, travel, military, *etc*), creating a separate DCM with the judgments of experts from these application areas may improve the model as well. With the application of this model to a new context, we demonstrate that it is practical to apply the type of aggregate expert knowledge captured in the DCM to new contexts in order to assist decision makers in making the complex choices at the technical frontier. In addition to possible benefits through the aggregation of expert knowledge, particularly intriguing may be the opportunity to quickly scale aggregate expert knowledge across a vast number of parts far greater

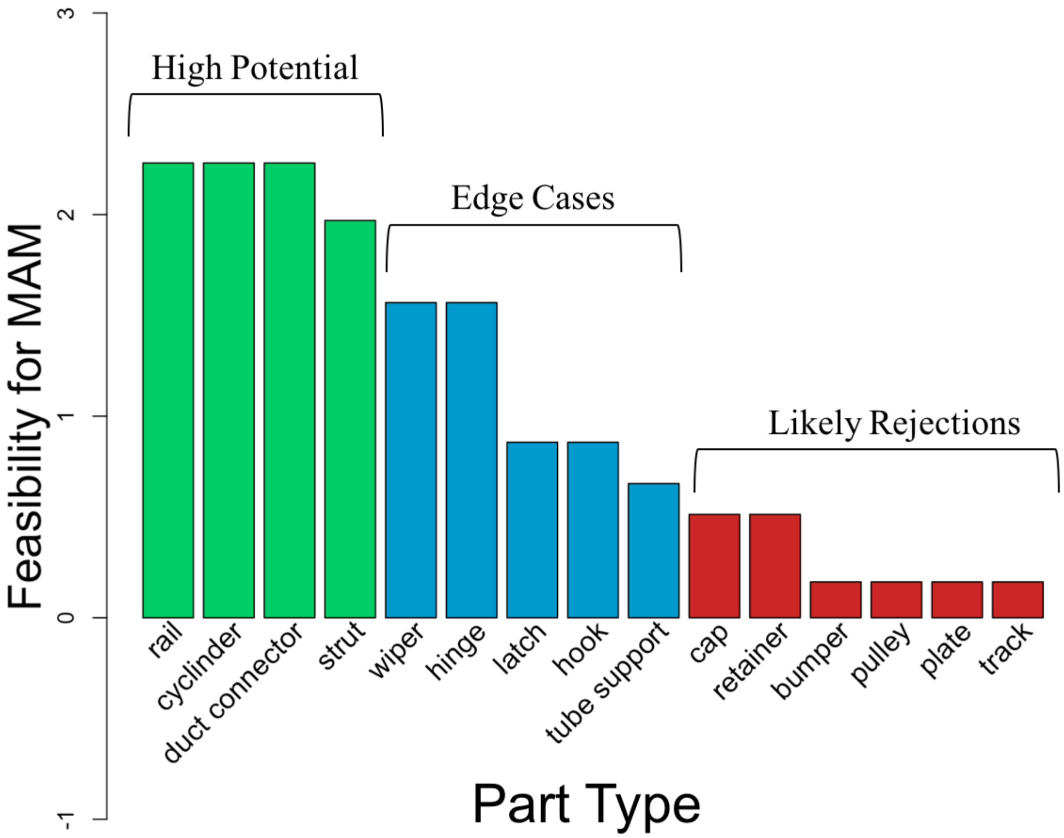


Figure 2.5: The modeled feasibility scores for the top parts as returned by the DCM applied to a set of new aircraft components.

than could easily be handled by available experts by hand, and generate a small number of part candidates for the new technology which could then be explored at greater length by experts (including for example, assessment, redesign, building, and testing).

## 2.4 Discussion and conclusions

Information asymmetries (Ackerloff, 1970; Hayek, 1945), insufficient human capital with knowledge in the new field (Crane, 1969), and the predominance of tacit knowledge(Bonnín Roca et al., 2017; Polanyi, 1966; Price, 1963) all create challenges for the diffusion and adoption of an emerging technology. Such emerging technologies rely heavily on small expert groups to drive the field, and yet research has also shown experts to be sub-optimal in prediction (Dawes & Corrigan, 1974) and to have difficulty in exactly the type of tasks required of early engineering challenges. Specifically experts struggle with decision rules when two alternatives appear equally viable (Marschak, 1959), time is a constraint (Busemeyer & Townsend, 1993), and preferences are context-sensitive (Tversky & Simonson, 1993; Tversky & Kahneman, 1986, 1981). Leveraging the example of metal additive manufacturing in aerospace, in which many of these factors that create challenges in the commercialization of a new technology are represented in particularly challenging forms, we seek to capture what expert decision-making looks like at the technological frontier and describe what opportunities may exist for interventions to accelerate such technologies' introduction.

Our survey seeks to codify first tacit and then explicit knowledge, with pairwise comparisons and attribute ratings, respectively. Differences between the part orders that emerge from our codification of tacit knowledge (our pairwise comparison data) versus from our discrete choice model – which draws on both our pairwise comparison data and our codification of explicit knowledge (the attribute ratings) – highlight potential differences between the tacit and explicit knowledge of our experts. Even when the sample set is restricted to only those experts who exhibit linear additive decision rules, the order returned by average pairwise comparison differs from the order predicted by applying the weights of the DCM to the explicit attributes (See Figure 2.6). The differences between the multidimensional scaling results and the discrete choice model outcomes suggest that there is something about the experts' decisions that is not yet captured in our survey. The MDS (which finds a lower dimensional spatial representation of the data) and DCM (which maps the choices to their explicit functional form) processes are not equivalent and therefore some degree of variation is expected. It is impossible to know whether any remaining missing knowledge is an attribute that we did not ask experts about or tacit knowledge that experts could not identify themselves. That said, when asked (both during pre-survey development and post-survey interviews), no experts felt there were major attributes missing, suggesting that the difference is likely based on tacit knowledge.

In the case of information asymmetries (Ackerloff, 1970; Hayek, 1945), the nature of our data makes it difficult to capture where differing knowledge, differing goals, or random variation in decision process may be causing variation in responses across experts. Our experts self-report industry, government and academic experience with many having all three. Analysis at the expertise level showed no discernible relation between judgments and expertise category. Due to the small number of experts in any single category this result is inconclusive. However, the large

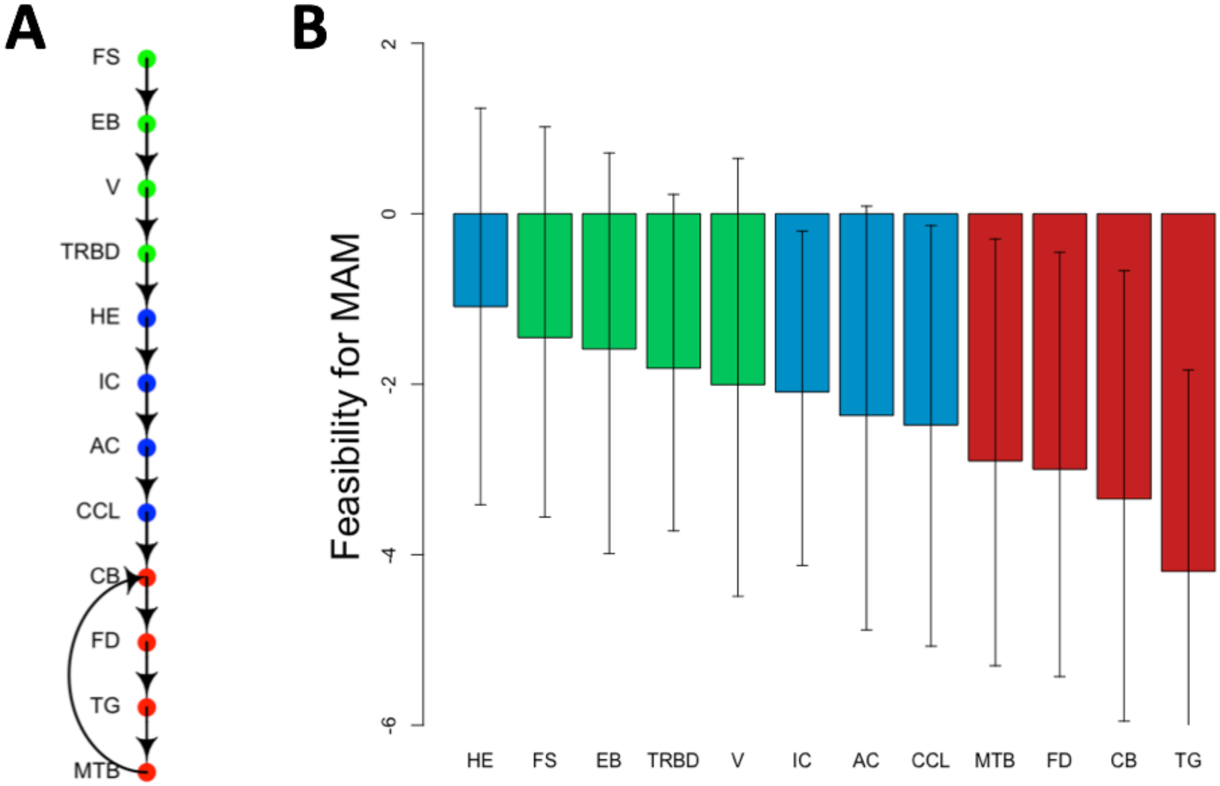


Figure 2.6: (A) The ordering of parts from a pairwise comparison of the 13 experts used in the DCM. (B) The part order found by applying the DCM to the expert choices and using a resampling bootstrap for 95% confidence intervals. There is a large degree of overlap in confidence intervals for the DCM order, but the pairwise comparison order has only one cycle in the lower section. Differing orders as well as degrees of confidence in the order are derived from each process, with the combination being the most informative. Color order from (A) is used in (B) with the heat exchanger (HE) being the only part out of its group placement.

disagreement across experts, including those within the same area of expertise, argues for differences being more likely a result of a disparity of knowledge rather than goals, as one would expect common goals within a given professional category. Even within the few experts who are consistent, all have different choice rules, leading to all unique judgments.

Inconsistency within and across experts, whether from decision process (Tversky, 1969), uncertainty, or randomness (Marschak, 1959), can be particularly challenging to overcome when the sample of experts is small, as is often the case with an emerging technology. Expert inconsistencies could be caused by two conflated factors. First, experts could be uncertain about what the parts themselves are or be uncertain about what the parts' technical specifications are (with the assumption that those details are required for any decision of feasibility to be made). With detailed, post-survey interviews with two-thirds of the respondents, we found no evidence that experts were confused about what the parts' functions were. This likely rules out the hypothesis that experts were confused about what parts they were shown. If, on the other hand, experts require more engineering detail to make consistent judgments of high-level feasibility for MAM, it emphasizes the need for methods such as ours to aid real world decision makers in sorting part portfolios. Second, it is possible that experts are certain about the stimulus or its details, but rather inconsistent in the application of their decision rule. Expert inconsistency also indicates methods such as the one outlined here, defining experts' judgment rules in order to apply them consistently across all part comparison scenarios, to move beyond qualitative assessments of the industry (Halal, Kull, & Leffmann, 1998; Jiang, Kleer, & Piller, 2017).

Strikingly, in our study, *each* of our 27 experts has a different judgment graph. Our simulations predict 10-15 expert responses are required for consistent decisions on aggregate. In our sample of 27 experts, no more than two came from any one organization, although several should have been within organization with similar goals. We suspect it is unlikely any one organization would have the requisite 10-15 experts. These results highlight the potential risks of single experts dominating the decision process within an organization and also the potential benefits of pooled expert knowledge for accelerating effective and efficient technology diffusion and adoption. Notably, these benefits of pooling were not assumed upfront in our analysis, but rather were seen in the data when averaging across expert responses. [3] These benefits may be even greater if expert knowledge is pooled across industry, academia, and government, as is done in our study. An additional issue of limited human capital is that there are likely too few experts to analyze all possible implementation opportunities. Given the potential fallibility of individual experts (assuming information asymmetries and random errors across experts and not that one of our experts is "correct" and the rest wrong), pooling and then scaling aggregate expert knowledge, such as that presented in the discrete choice model, could have benefits for accelerating effective and efficient technology implementation. This possibility is tentatively explored in our application proof-of-concept.

From a theoretical standpoint, our results build new understanding of how individual judgment structure relates to group judgments. First, we find that the majority of experts are internally inconsistent (intransitive in their judgments), and yet when we aggregate across experts, a judgment order emerges that is nearly consistent. Intriguingly, this finding is an inversion of the Condorcet paradox (de Condorcet & Diannyére, 1797), in which individually consistent decision-makers have inconsistent judgments when aggregated. Second, we find that individual experts who are far from the aggregate of the experts' knowledge are more inconsistent (they have larger MFES). The only experts who are consistent are those whose responses suggest they are using simplifying rules (either single attribute criteria or decision trees). However, this consistency and these simplifying rules do not guarantee agreement as all seven of our consistent experts have different judgment orders and different choice processes. While prior work has found that people use simplifying heuristics that can ensure rationality (Brandstätter, Gigerenzer, & Hertwig, 2006), to our knowledge this is the first research to show that those experts using simplifying heuristics may also be closer to the aggregate of expert knowledge when all experts disagree (where it remains unknown if that aggregate is correct or incorrect). In summary, this is first work to show that (a) multiple consistent individuals who use simplifying heuristics can have inconsistent aggregate judgments, (b) multiple inconsistent individuals who use more complex decision rules can have

[3] By having independent survey responses from experts across the field of metal additive manufacturing for aerospace, we have mitigated the risk of group think in our data. Table 1 shows that our experts have a large variety of backgrounds; indeed several experts have worked across multiple sectors. One of the findings of our study is that there is a significant diversity in the extent to which experts' choices are in line with the group average (see the Hamming Distance, HD, on the x axis of Figure 2(A)). This suggests that there isn't strong evidence of group think among our respondents. That being said, there is always a chance that collectively experts are not on the most valuable path forward. At the technical frontier, this is challenging to verify given the difficulty of determining ground truth comparisons.

consistent aggregate judgments. Together, (c) these results suggest that individual inconsistency may facilitate coherent group decision-making. Understanding when aggregate consistency is a function of all experts agreeing or, as in our case, experts disagreeing may be critical for decision making, as aggregation could both provide new insights but also mask important disagreement and uncertainty in complex decision environments.

As suggested above, in our sample, some experts are likely making more accurate predictions than others. To improve the prediction process, ideally one would over-weight those who are more accurate in their evaluations when combining experts. There are a variety of methods for weighting experts, but most of them rely on measuring against known calibration variables (Colson & Cooke, 2017; R. M. Cooke, 1988). At the technical frontier, where calibration is difficult to define and expensive to validate, additional techniques are needed. A recent method developed by Prelec, *et al.* in 2017 (Prelec, Seung, & McCoy, 2017) and tested by Lee *et al.* in 2018 (Lee, Danileiko, & Vi, 2018), known as the *surprisingly popular method*, does not require calibration variables but instead relies on additional confidence measures to add nuance to the pairwise comparison responses. Combining that method with a physical production and testing regime in the future would offer new insights into not only expert judgment and decision-making, but also the implications of decisions as they drive the adoption of new materials and processes (here, metal additive manufacturing in the aerospace community) in the wider economy.

Previous research on contexts where knowledge, methods, and problems are unknown (the scientific frontier) (Dunbar, 1997), and also where knowledge, and methods are known, but problems are new (McComb et al., 2017), find that, like lay people (Tversky & Kahneman, 1974), experts employ various heuristic strategies to substitute difficult decision processes for simpler ones. Focusing on an emerging technology which offers an extreme example of the tacit knowledge, information asymmetries and insufficient human capital that can delay new technology adoption, we codify expert decision-making at the technological frontier. We find that no two experts have the same judgments and almost all have internal inconsistencies, but when their knowledge is aggregated a clear and mostly consistent decision-rule emerges. These results suggest that capturing, pooling, and then scaling that aggregate expert knowledge may have potential for accelerating commercialization of new materials and processes.

# 3

# Bounds on the judgmental accuracy of complete pairwise comparison rankings in the absence of outcome information

## 3.1 Introduction

Judgments from subject-matter experts can be used for prediction, either by combining those judgments with statistical models, or by using judgments directly when data are expensive or difficult to collect. For example, analysts might ask political science professors to rank presidential candidates based on their likelihood of winning the next election, or critical care physicians might be asked to determine which patients are more likely to survive a viral infection if provided specialized medical treatment. In each case, decisions must be made before relevant data are available, requiring the use of expert judgment. Although subject-matter experts have extensive knowledge about the topic, the accuracy of expert judgments can vary significantly from expert to expert (Funk, Davis, Vaishnav, Dewitt, & Fuchs, 2020). To address this variability, analysts might collect data to validate the expert, yet in many cases validation data are unavailable (the election hasn't happened yet; patients need treatment immediately). Ideally a method could be used to validate the quality of an expert's judgments at the time of prediction, before critical outcomes occur, rather than from past predictions that may not be applicable to a new setting.

When experts are asked to make judgments in the form of rankings over the likelihood of events, transitivity might be considered a measure of response quality. For pairwise comparisons between the subsets of three alternatives $\{A, B, C\}$, a transitive ranking $\succ$ on the set requires that if A $\succ$ B (A is judged to be more likely than B), and B $\succ$ C (B is judged to be more likely than C), then A $\succ$ C (A is judged to be more likely than C). Behavioral decision researchers interpret intransitivity as revealing uncertainty about the judge's knowledge or preferences, or alternatively that judges are using heuristics to simplify complex tasks. For example, the lexicographic semiorder (Tversky, 1969) is a simple decision rule that can lead to intransitive choices.

Intransitivites are sometimes seen as indicators of poor judgment and reasons to exclude judges from a prediction task. For a single expert, intransitivity makes it impossible to create a complete ranking from the expert's predictions, an intuitively appealing reason for exclusion from aggregation with other experts. However, there is some evidence that use of heuristics can help produce more accurate rankings (Arkes, Gigerenzer, & Hertwig, 2016; Czerlinski, Gigerenzer, & Goldstein, 1999), even though use of heuristics risks intransitivity. Here we suggest that intransitivity should not be seen as a definitive factor for rejecting an expert, nor an informationless feature unrelated to prediction accuracy. Rather, by leveraging the degree of an expert's intransitivity to bound their possible accuracy, we describe the tradeoff between seeking perfectly transitive experts who may be very wrong, and intransitive experts who can't be completely right. Specifically, we prove that the degree of intransitivity of an expert's judgments bounds their potential accuracy, making it possible to assess *potential* prediction success without any outcome information.

We consider the problem of determining which of two judges is more accurate at making predictions on a task that does not have ground truth data available. By only observing the predictions of the two judges, without also observing whether their predictions are right or wrong, it is impossible to tell which one is more accurate. The range of possible accuracy for each of the judges is, of course, the closed unit-interval $[0, 1]$, no matter how many judgments they make, unless some outcomes are also observed. Consider a second case where the judge assigns a score $s_i \in \mathbb{R}$ meant to be proportional to the likelihood of event $i$ out of $n$ total events. The judge considers the event $i$ as being more likely than event $j$ if and only if $s_i > s_j$. A judge assigning

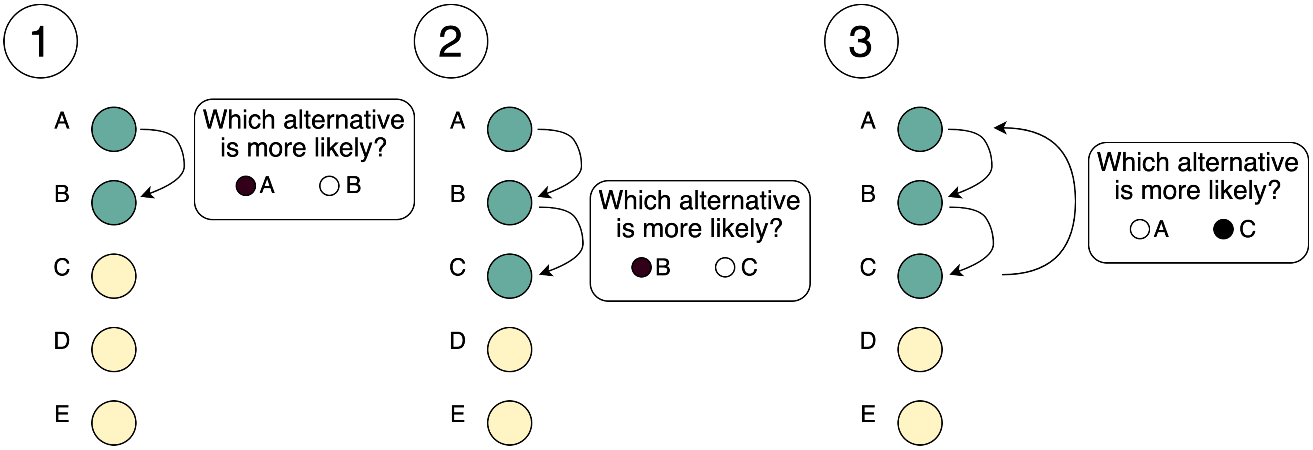

1
A
B
C
D
E
Which alternative is more likely?
● A ○ B
2
A
B
C
D
E
Which alternative is more likely?
● B ○ C
3
A
B
C
D
E
Which alternative is more likely?
○ A ● C

Figure 3.1: The above is an example of how a cycle is uncovered in a person's choice structure and how that cycle is represented graphically. 1 - Alternative A is chosen over Alternative B (represented by a directed edge on the graph). 2 - Alternative B is chosen over Alternative C. 3 - Alternative C is chosen over Alternative A.

scores to each event implicitly provides a ranking over the events. The judge's skill can be calculated using the area under the ROC curve (AUROC), which is just the proportion of rankings that are correct among all pairs of the $n$ events. The equivalence between accuracy and AUROC for two alternative forced choice tasks is sometimes called the two alternative forced choice theorem (Bamber, 1975). Because the judgments were elicited as scores, they must form a total order. However, without any information about the ground truth of the outcomes, the judge may get the order completely correct, or completely incorrect.

In this paper we focus on a third case shown in Figure **??**, where the judge chooses which of two events $i, j \in \{1, 2, \ldots, n\}$ is more likely for all pairs of events. In this case it is possible for a judge to provide an intransitive (cyclic) set of judgments (shown in Figure **??**, sub-plot 3). We show that in the elicitation of pairwise comparisons about the likelihood of events, it is possible to determine which judge has the better *possible maximum (or minimum)* skill, as defined by accuracy or the AUROC. That maximum or minimum possible skill is defined exactly by the cardinality of the *minimum feedback edge set* (Ali, Cook, & Kress, 1986) of a *tournament graph* (Moon, 2015), or the minimum number of edges in a *complete directed graph* that need to be reversed to make it acyclic. These bounds can be constructed without any information on the ground truth of the predictions.

## 3.2 Proof of the relationship between AUROC and the minimum feedback edge set of a tournament graph

In this section we prove the following:

**Theorem 3.2.1.** *Let $h \in T_n$ be a tournament graph with $n$ vertices and $v$ edges (where $v = \binom{n}{2}$). Let $g \in G_n$ be an acyclic tournament graph on $n$ vertices. Call $AUROC(h,g) : T_n \times G_n \to [0,1]$ the* AUROC function *that maps pairs of tournament graphs and acyclic tournament graphs on $n$ vertices to $[0,1]$, and $|MFES(h)| : T_n \to \mathbb{N}$ is the* minimum feedback edge set cardinality function *that maps tournament graphs to the natural numbers. Then,*

$$\max_{g \in G_n}(AUROC(h,g)) = 1 - \frac{|MFES(h)|}{2v} \tag{3.1}$$

### 3.2.1 Preliminaries

Define a *tournament graph* $h \in T_n = \{\text{all tournaments with } n \text{ vertices}\}$ as any $n \times n$ adjacency matrix that satisfies the following three conditions: 1) $h_{ij} \in \{0,1\}$ for all $i, j \in \mathbb{E}$ where $\mathbb{E}$ is the edge set of $h$, 2) $h_{ii} = 1$ for all $i = j \in \mathbb{E}$, and 3) $h_{ij} + h_{ji} = 1$ for all $i \neq j$. An *acyclic* tournament graph $g \in G_n$ is a tournament graph with the additional restriction that there is a permutation matrix $\rho$ in the set of $n \times n$ permutation matrices $\mathbb{P}$ such that $\rho g$ is upper triangular. Let $K_{hg} = \sum_{i=1}^{n}\sum_{j=1}^{n}(h_{ij} - g_{ij})^2$ be the number of edges that are not equal in $h$ and $g$. Then,

$$\begin{aligned}
K_{hg} &= \sum_{i=1}^{n}\sum_{j=1}^{n}(h_{ij} - g_{ij})^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}(h_{ij}^2 - 2h_{ij}g_{ij} + g_{ij}^2) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij} - 2\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}g_{ij} && \text{because } h_{ij}, g_{ij} \in \{0,1\} \\
&= 2v - L_{hg},
\end{aligned}$$

where $v = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} = \frac{n(n+1)}{2}$ is the total number of edges in a tournament graph of $n$ vertices. $L_{hg} = 2 \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} g_{ij}$ is the number of edge agreements between $h$ and $g$ because $h_{ij} g_{ij} = 1$ only when $h_{ij} = 1$ and $g_{ij} = 1$, and when both are zero it must be the case that their complements are both 1, so multiplying by two captures the case when both are zero.

Let the graph $g^*$ be the (possibly non-unique) acyclic graph that is closest to a tournament $h$ in *squared distance*:

$$g^* =_{g \in G_n} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} (h_{ij} - g_{ij})^2 \right) \tag{3.2}$$

The graph $g^*$ can be thought of as the acyclic graph that requires the fewest edge reversals to make $g^*$ equal to $h$. The set of mismatched edges between $g^*$ and $h$ forms the *minimum feedback edge set*, with cardinality:

$$K_h^* = \min_{g \in G_n}(K_{hg}) = \sum_{i=1}^{n} \sum_{j=1}^{n} (h_{ij} - g_{ij}^*)^2 \tag{3.3}$$

Define the *area under the ROC curve* (AUROC) between tournament graph $h$ and acyclic graph $g$ in the following way. Suppose we have $n$ objects that form a total order. Those $n$ objects can then be represented as a graph $g \in G_n$ such that $g_{ij} = 1$ if $i$ is ranked higher than $j$ and is zero otherwise. An edge in graph $h$ is *correct* if $h_{ij} = g_{ij}$, and *incorrect* otherwise. AUROC is then $P(h_{ij} = g_{ij})$, or the proportion of agreeing edges between the two graphs. Thus, AUROC can be expressed as $AUROC(h, g) = \frac{L_{hg}}{2v}$.

## 3.3 Result

We now prove Theorem 1, that the maximum possible AUROC for a tournament graph is related to the cardinality of the minimum feedback edge set in the following way:

$$\max_{g \in G_n}(AUROC(h, g)) = 1 - \frac{|MFES(h)|}{2v}$$

*Proof*: From above, we have $K_{hg} = 2v - L_{hg}$, giving

$$\begin{aligned} K_h^* &= \min_{g \in G_n}(K_{hg}) \\ &= \min_{g \in G_n}(2v - L_{hg}) \\ &= 2v + \min_{g \in G_n}(-L_{hg}) && \text{because } v \text{ is constant} \\ &= 2v - \max_{g \in G_n}(L_{hg}). \end{aligned}$$

AUROC can be expressed as $AUROC(h, g) = \frac{L_{hg}}{2v}$, giving

$$\max_{g \in G_n}(AUROC(h, g)) = \frac{\max_{g \in G_n}(L_{hg})}{2v} \tag{3.4}$$

Rearranging gives $2v \cdot \max_{g \in G_n}(AUROC(h, g)) = \max_{g \in G_n}(L_{hg})$. Substituting this into equation 4 gives $K_h^* = 2v - 2v \max_{g \in G_n}(AUROC(h, g))$. Solving for $\max_{g \in G_n}(AUROC(h, g))$ gives the result,

$$\max_{g \in G_n}(AUROC(h, g)) = 1 - \frac{|MFES(h)|}{2v} \tag{3.5}$$

where $|MFES(H)| = K_h^*$ is the cardinality of the minimum feedback edge set of $h$. □

As an example, if we have a graph with 10 vertices (10 alternatives in a complete pairwise comparison task), the possible MFES sizes range from 1 to 16. Since $v = 45$ when $n = 10$, this gives an max(AUROC) range of approximately .99 to .82 (respectively).

## 3.4 Simulation experiments using the bound

In this section we explore, through simulations, the conditions under which knowledge of the cardinality of the minimum feedback edge set can improve AUROC bounds, aggregation, and search during judgment elicitation. First, we examine single-expert bounds by simulating graphs with 10 or fewer alternatives, characterizing the bounds by cardinality of the minimum feedback edge set. Second, we evaluate strategies that use only the cardinality of the minimum feedback edge set for aggregation. Third, we examine how the cardinality of the minimum feedback edge set can be used to improve judgment elicitation protocols.

### 3.4.1 Bounds

**Motivation**

We begin by simulating all possible tournament graphs with sizes of five to ten vertices. Such graphs might be elicited from asking political science experts about the prospects of primary presidential candidates, asking virologist about the relative potential of pathways to treat a novel virus, or asking head coaches of a professional sports league about the future career success of the top picks from an upcoming draft. In areas like these, where outcome probabilities are valuable but prediction accuracy is difficult to measure, the MFES of expert responses may provide some insight into experts' calibration to the task.

**Method**

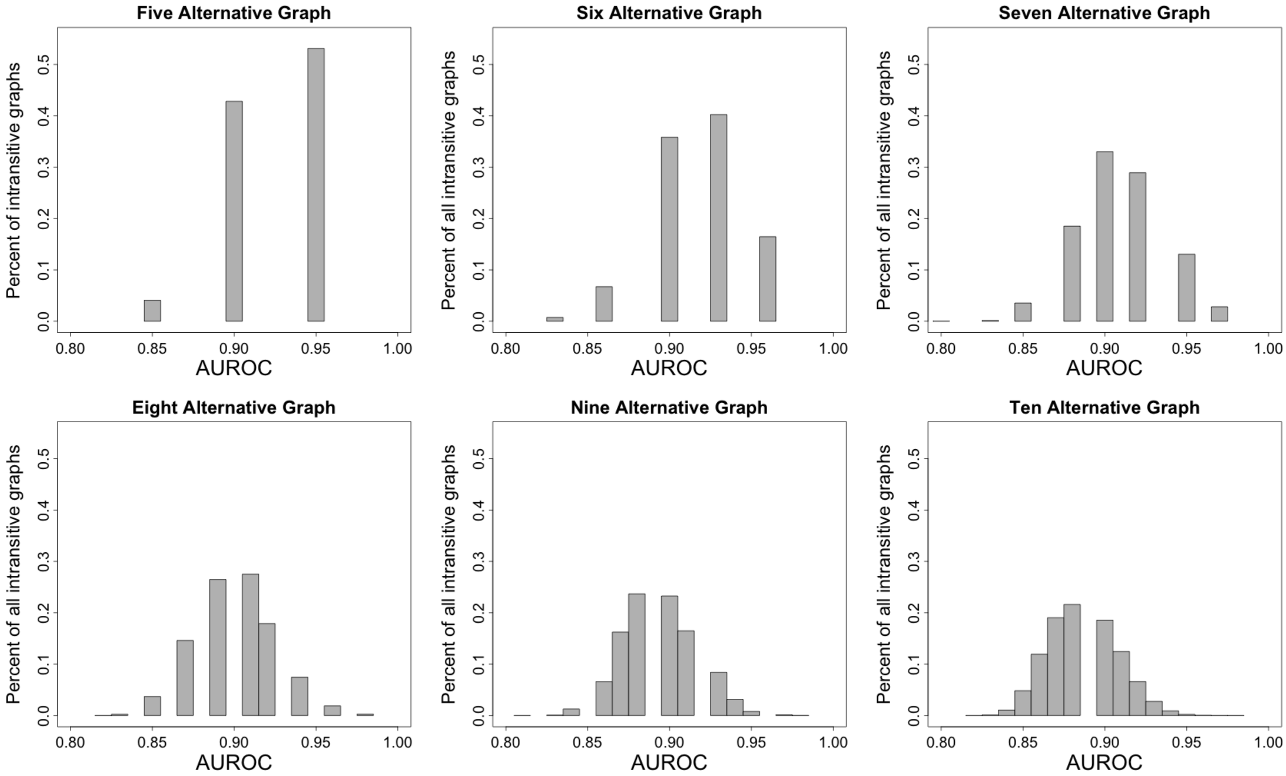
Tournament graphs are simulated by randomly generating all $h_{ij}$ pairwise comparisons of $n$ alternatives from independent Bernoulli distributions ($p = 0.5$) to create adjacency matrices. These matrices yield random graph structures and random (potentially cyclic) rankings sampled from all possible tournament graphs of size $n = 5$ to $n = 10$ vertices. Graphs of $n < 5$ are trivial with respect to MFES as $n = 2$ cannot have intransitivity and $n = 3, 4$ have only one possible MFES size of 1. With large enough samples it is possible to enumerate all graph structures and rankings. For each random graph, the MFES is calculated using the kwiksort algorithm (Ailon, Charikar, & Newman, 2008). The simulation is repeated 10 million times for each $n$ to provide a sample of how MFES scales with small $n$ (the approximate range of a plausible complete pairwise comparison task). These graphs represent the space of all possible graphs from complete pairwise comparison tasks, which may not be equally likely in practice. Given that our example case is of evaluating prediction results *prior* to seeing the outcomes, generating the full space of graphs is the most general approach.

**Result**

Figure 3.2 shows the possible AUROC values for a complete pairwise comparison task given the set of possible MFES sizes from a graph of $n = 5$ to $n = 10$ alternatives. At $n = 5$, there are relatively few possible MFES states and a left skew to the AUROC distribution. As $n$ increases, the distributions remain in the $0.8 - 0.99$ range, but the mode of the distributions shifts left slightly. While we focus on the upper bound of the AUROC, the lower bound is raised proportionally for each possible MFES size. As graph sizes increase, the max(MFES) also increases, but at a slower rate than the overall graph (de la Vega, 1983). The exact computation of the MFES is NP-hard, but Figure 3.2 suggests that for small $n$, as found in experimental choice settings, AUROC for cyclic responses will be bound in the range of 85% to 95% (Alon, 2006; Broomell & Bhatia, 2014). From the expert decision making literature, we may expect both experts and statistical models of expert decisions to fall in this range of accuracy (Dawes et al., 1989). Depending on the exact size of the prediction task then, this bound may be able to sort which experts can do better than a model before a benchmark is known. In tasks where statistical models reach higher accuracy levels, this method will likely not be a tight enough bound on experts to guarantee improvement over the statistical model (Rajkomar et al., 2018).

Caution must be taken in this approach as well considering the MFES and total number of pairwise comparisons do not grow at equal rates with the number of alternatives. The number of comparisons is always $\frac{n(n-1)}{2}$. The MFES has no closed form solution, but grows much more slowly for small n and at large n is on the order of $n^{\frac{3}{2}}$. As the graph size increases, the maximum possible MFES therefore will grow more slowly than the size of the graph and the bounds on accuracy become less restrictive. This feature forces one to think about this bound in terms of the specific task rather than in absolute terms, as adding additional alternatives to a comparison can provide a different bound without changes to the MFES size of the respondent.

Unfortunately, calculating the MFES is an NP-hard problem, making the measurement of how the MFES and our bound change with very large $n$ computationally intractable (Alon, 2006). Spencer (1980) and de la Vega (1983) proved bounds on the MFES, but more refinement of these limits will be necessary to understand the behavior of accuracy at larger $n$ (Spencer, 1980; de la



Five Alternative Graph

Six Alternative Graph

Seven Alternative Graph

Eight Alternative Graph

Nine Alternative Graph

Ten Alternative Graph

Figure 3.2: As graphs increase in size (shown from $n = 5$ to $n = 10$ here) the possible accuracy states for intransitive experts increase in number but also change in distribution.

Vega, 1983). It should be noted, however, that asymptotic bounds are useful theoretically, but a complete pairwise comparison task on large $n$ is likely infeasible in the discrete choice setting.

### 3.4.2 Aggregation

#### Motivation

Aggregating predictions helps leverage the wisdom of the crowd to improve over a single response. There is an ongoing discussion in the aggregation literature on how or if to weight experts when aggregating their responses (Genest, Zidek, et al., 1986; Prelec et al., 2017; R. Cooke et al., 1991; Colson & Cooke, 2017). While some of these methods aim to weight experts by their accuracy on previous tasks, there is not always an appropriate correlate prediction with which to benchmark each expert. Given that intransitivity in a prediction task is traditionally considered a violation of rational behavior, we examine if it can be used to help weight experts without requiring a benchmark. Here we explore the possibility of using the MFES and resulting maximum or minimum possible AUROC to inform aggregation of responses.

#### Method

Adjacency graphs are created using the methods described in Section 3.4.1. MFES is calculated again using the kwiksort algorithm and binned by MFES for grouping (Ailon et al., 2008). Aggregation is done by averaging the adjacency matrices and rounding the resulting average matrix to a binary adjacency matrix. A randomly generated *correct* order is used as the baseline to measure the aggregated result's accuracy.

#### Result

We find that without prior information to benchmark predictions, it is not possible to improve AUROC by choosing an aggregation method based on the degree of inconsistency of the pool of respondents. Figure 3.3 top left shows no change in AUROC with simple averaging using all acyclic experts or grouping experts by MFES size. This suggests that in a situation where all orders are equally likely (or we have no *a priori* information to say otherwise), knowing the structure of responses provides no benefit in expected accuracy.

The variance of the accuracy is affected by structure, however. Figure 3.3 top left shows the simulated AUC decreases as group size increases, as one would expect. Variance also decreases in groups composed of members with larger MFES. The top right and bottom two panels show that as we increase the threshold of accuracy from 70% to 90%, their is both a reduction in accuracy and reduction in variance with the inclusion of cyclic respondents. If we presume the respondent pool is more accurate (as is likely the case with experts), the average accuracy is reduced with inclusion of cyclic experts, and the benefits of the reduction in variance are limited compared to the random sampling. If one is running a prediction task aggregating expert responses, it may be beneficial, from a time and cost perspective, to filter out those experts who have large MFES as early as possible.

### 3.4.3 Search

#### Motivation

Experts' time is valuable. A good elicitation protocol balances gaining as much information as possible against minimizing the time required for the expert. A practitioner can use our bound to filter experts during an elicitation protocol, ending a process early if a MFES beyond the desired bound is revealed or using the intransitivity from the MFES as a point of discussion mid-elicitation, rather than after the elicitation is complete. This will be most useful when a benchmark prediction model exists for the area of interest whether that be from a statistical model (Langlotz et al., 2019), historical case study (Dawes, 1979), or a prediction market or forecasting tournament (Mellers et al., 2014).

#### Method

Consider a complete pairwise comparison task of five elements. There are 10 total comparisons (graph edges) required to compare each element against every other. In a traditional elicitation process, an expert is asked to answer all 10 comparisons in random order to determine their ranking of the five elements. With the bounds proven above, we posit two alternative elicitation

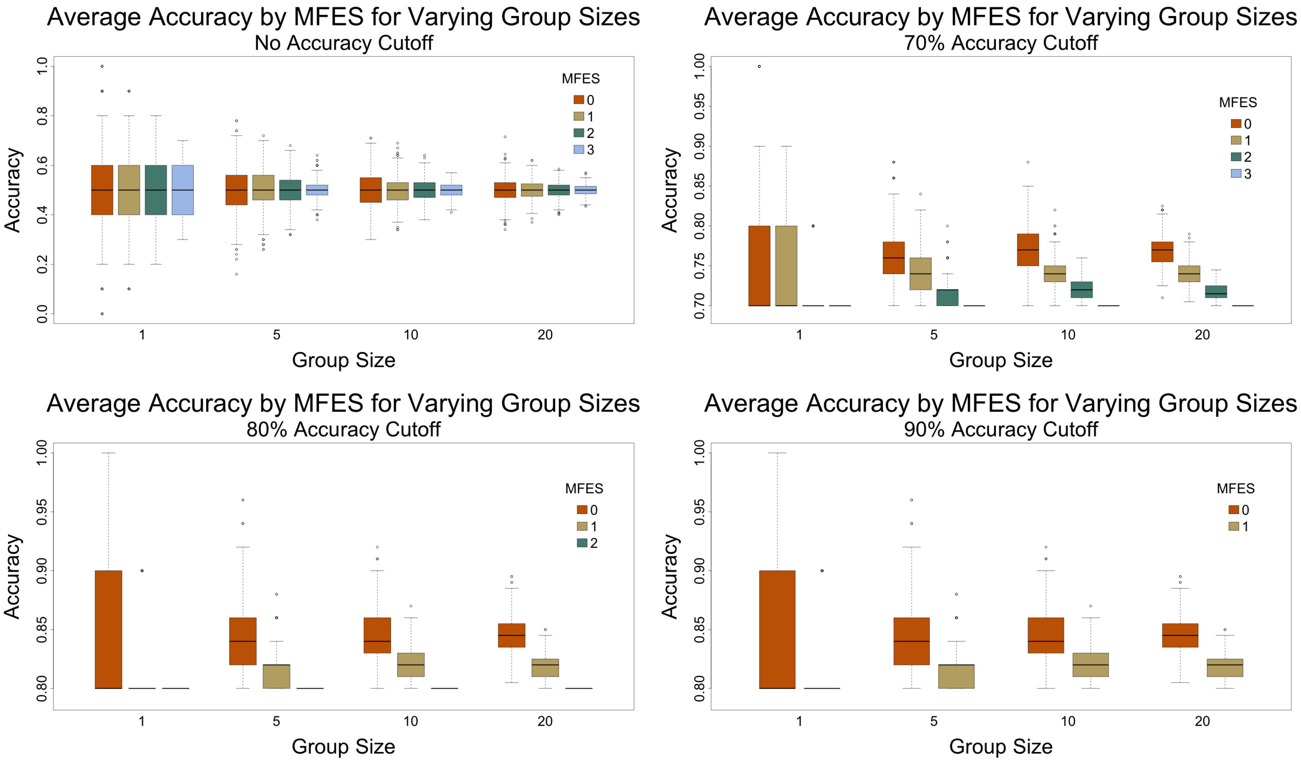


Figure 3.3: Four simulation scenarios show the effect of group size and MFES size in aggregation with differing expectations of the level of accuracy within the pool of respondents. Top Left: With random sampling from all possible responses (as may be the case when the expertise level within the pool is unknown), including cyclic experts in aggregation reduces the variance of prediction accuracy. Top Right: If we assume a minimum accuracy of 70% within the pool, including cyclic experts still reduces variance, but it also reduces the average accuracy. It should be noted that the only possible graph with MFES = 3 in this case has *exactly* a 70%, hence the compressed distribution. Bottom Left: The possible MFES sizes are reduced with an 80% accuracy threshold. A slight variance reduction with a decreased accuracy are seen here again. Bottom Right: Similar behavior is shown with a 90% accuracy threshold: reduced possible graph structures, decreased average accuracy with cyclic graphs, and decreased variance with cyclic graphs.

protocols (as compared to a random presentation of comparisons). 1) Random Search: ask comparisons in a random order and check after each question whether the expert has fallen below the desired threshold of accuracy due to revealed cycles (knowing that only when a complete subgraph of the respondents preferences has been revealed can a MFES determination be made). 2) Strategic Search: to increase the efficiency further, ask questions in a semi-random process by randomly adding the next alternative of comparison but comparing it to all previous alternatives before moving on in the elicitation.

To test the impact of these search methods we simulate a search process across 10,000 random iterations of each possible cyclic 5-graph. Each search starts with the same randomly generated pair of alternatives and then proceeds through the Random Algorithm 1 and the Strategic Algorithm 2 (graphical representations can be found in Figure 3.4). The cutoff for the simulation is an AUC of 95% (The simulation ends if a structure is revealed that dictates a maximum AUC of less than 95% from the above proven bound).

**Algorithm 1** Random Search

```
Randomly select the first alternative
Select the next alternative to form the first pairwise comparison
while Asked comparisons < total comparisons do
    if A subgraph is completed then
        if The MFES of the subgraph > 0 then
            if The MFES > desired bound then
                End elicitation
            end if
        end if
    else
        Select an unasked pairwise comparison
    end if
end while
```

### Result

Figure 3.5 shows the number of questions saved as a function of specific cyclic structures and as a function of the general size of MFES for a 5-graph. Letters A through K represent the 11 possible cyclic structures of a complete pairwise comparison graph with 5 vertices (outlined visually in Moon (Moon, 2015)). Simulations show an average reduction in required questions of 1 to 5.5 out of

**Algorithm 2** Strategic Search

```
Randomly select the first alternative
Randomly select the next alternative to form the first pairwise comparison
while Asked comparisons < total comparisons do
    Select the next alternative randomly and present all comparisons that include this alternative
    and the previous alternatives, completing a subgraph between the alternatives
    if The MFES of the subgraph > 0 then
        if The MFES > desired bound then
            End elicitation
        end if
    else
        Select an unexamined alternative and compare it to all previous alternatives to create the
        next larger subgraph
    end if
end while
```

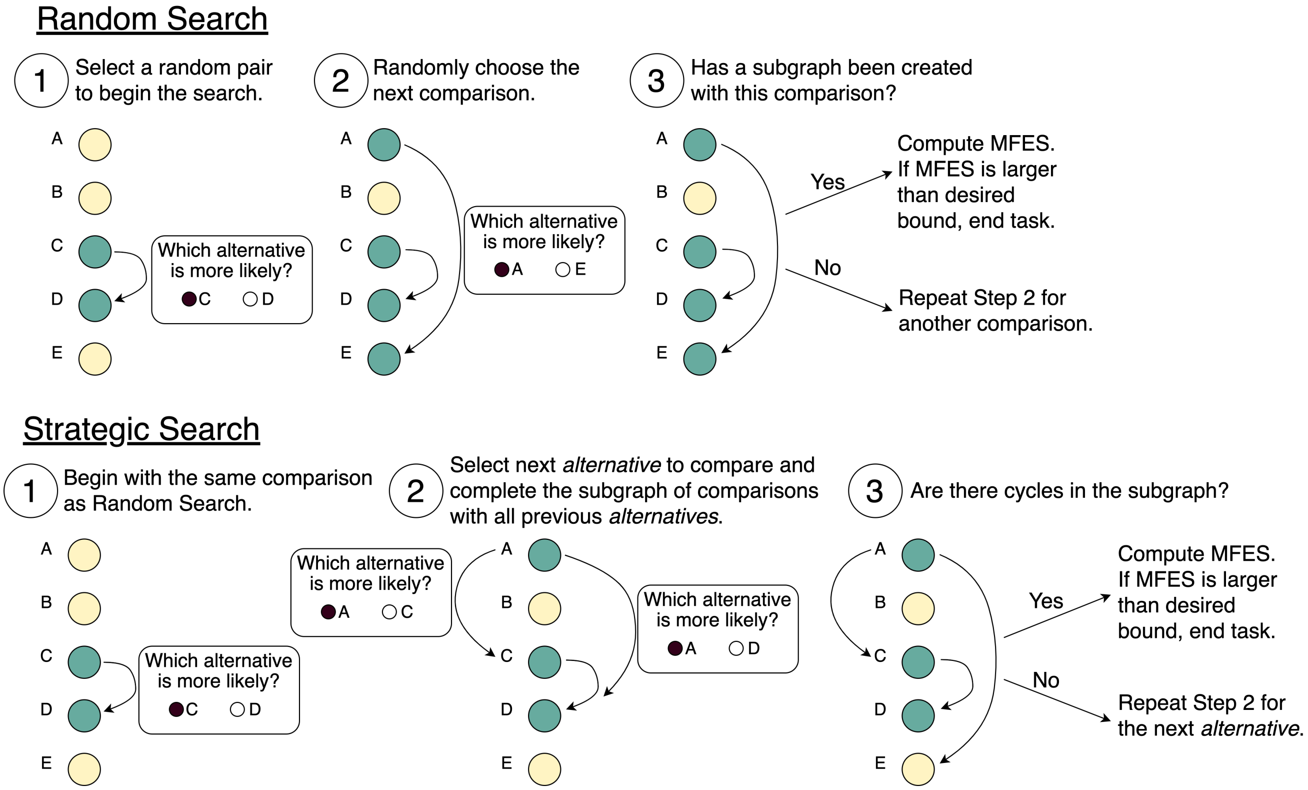

Random Search
1 Select a random pair to begin the search.
A
B
C
D
E
Which alternative is more likely?
● C ○ D
2 Randomly choose the next comparison.
A
B
C
D
E
Which alternative is more likely?
● A ○ E
3 Has a subgraph been created with this comparison?
A
B
C
D
E
Yes
Compute MFES. If MFES is larger than desired bound, end task.
No
Repeat Step 2 for another comparison.
Strategic Search
1 Begin with the same comparison as Random Search.
A
B
C
D
E
Which alternative is more likely?
● C ○ D
2 Select next *alternative* to compare and complete the subgraph of comparisons with all previous *alternatives*.
Which alternative is more likely?
● A ○ C
A
B
C
D
E
Which alternative is more likely?
● A ○ D
3 Are there cycles in the subgraph?
A
B
C
D
E
Yes
Compute MFES. If MFES is larger than desired bound, end task.
No
Repeat Step 2 for the next *alternative*.

Figure 3.4: *Random Search* and *Strategic Search* offer two approaches to using the proof bound to increase the efficiency of an elicitation process.

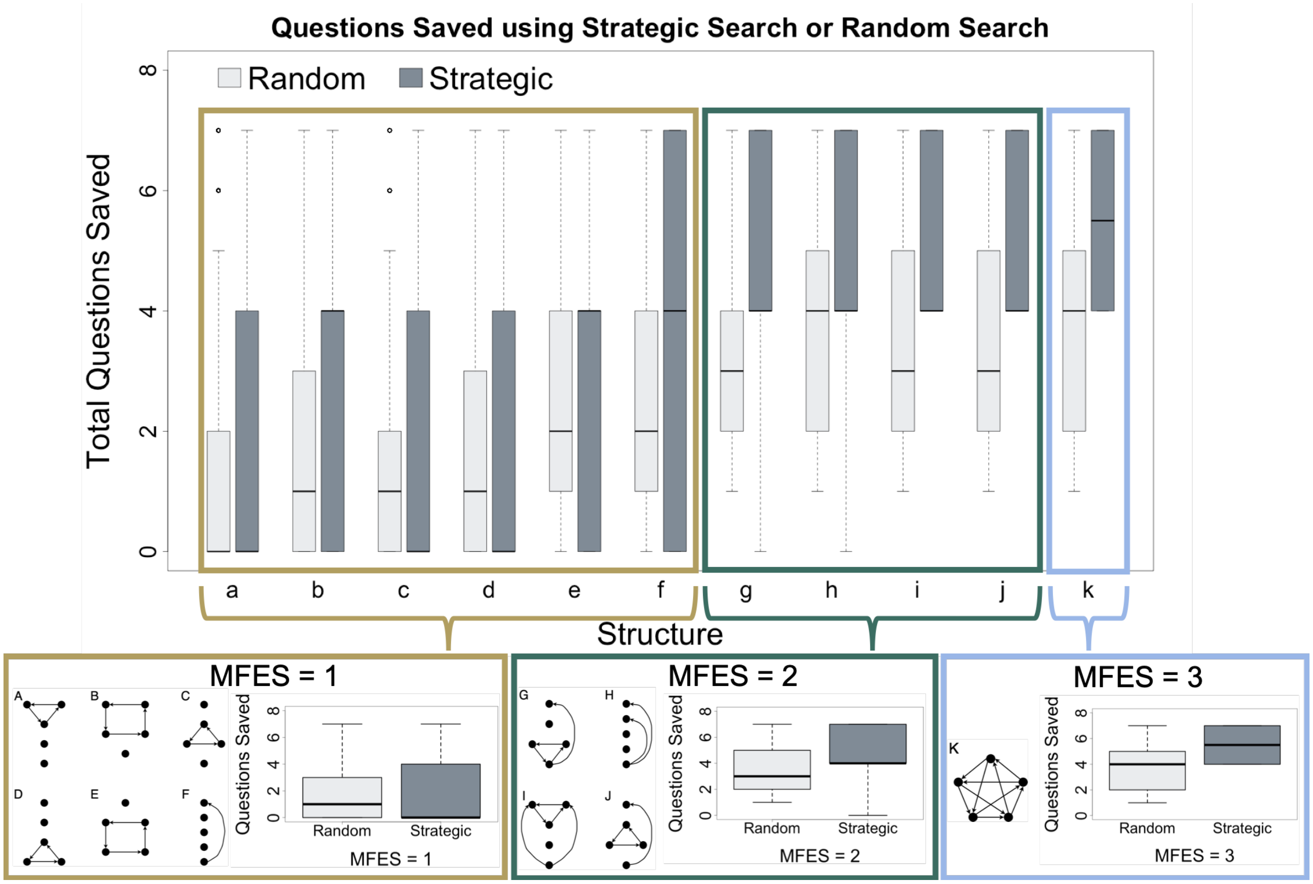


Figure 3.5: Top - For respondents that exhibit any of the 11 possible cyclic structures on a 5-graph (Moon, 2015), there is a decrease in the average number of comparisons required before a cycle is revealed as compared to a complete pairwise comparison. Below - The specific structures and averages for those MFES sizes show an increasing benefit of using the Strategic Search over Random Search or the traditional complete pairwise comparison procedure.

10 possible questions (10 to 55 percent). The potential questions saved increase with the size of the MFES since complex cyclic structures have more search paths that will reveal cycles. The Strategic Algorithm offers a greater average savings across all structures, though the algorithms each have the same range. This is because the Random Algorithm follows the same search path as the Strategic Algorithm in the best case. The benefits of using the algorithms will change depending on the desired accuracy cutoff (a lower accuracy cutoff will save fewer questions as some cyclic structures may be allowed) and number of alternatives (more alternatives in the prediction task will change the possible MFES sizes and accuracy bounds as discussed in 3.4.1).

## 3.5 Discussion

In many fields, expert judgment is used in place of ground truth data when the outcome of interest is expensive to obtain or only exists in the future (R. M. Cooke, 1988; Mandel & Barnes, 2014; Abdollahi et al., 2018; Morgan, 2014). One method for eliciting these judgments is through a prediction task of pairwise comparisons between alternatives of interest. These pairwise comparisons create a preference graph that describes the expert's rank of the alternatives. Transitivity of the resulting preference graph is often used as a metric of quality of a judgment, as transitivity "is the cornerstone of normative and descriptive decision theories" (Tversky, 1969). This assessment metric is not universal, however, as many decision models have the possibility of inducing intransitivity (Tversky, 1969; Gigerenzer & Goldstein, 1996). Our work highlights that there is information gained from thinking of transitivity on a spectrum, rather than a binary inclusion/exclusion criteria. We show that there is a relationship between the degree of intransitivity of a set of choices (minimum feedback edge set of a preference graph) and the potential accuracy of the respondent (the maximum possible area under the receiver operating curve (AUROC)).

There are also consequences when aggregating across respondents. Dropping intransitive experts, due to their supposed violations of rationality, increases the variance of the average response. This effect is countered by an increased average in groups with a high accuracy threshold, however. Disagreement amongst experts is an important feature of most aggregation methods for prediction (Prelec et al., 2017; Navajas, Niella, Garbulsky, Bahrami, & Sigman, 2018). Our result suggests that in settings without an accuracy threshold (or where one does not expect the average accuracy to be high), including intransitive responses will reduce the variance of the aggregate. In expert settings however, this variance reduction comes at the cost of reduced accuracy. Contextual information will also play a role in determining when to apply this bound as a tool for choice tasks. From a practical perspective, variance reduction may be important in risk-averse areas where limiting the worst responses is more valuable than perfect accuracy. Some areas of medical decision making have well established metrics for judgment that can improve the differentiation between judgment and choice heterogeneity (Tsuchiya et al., 2002). When attributes of alternatives in a task are well-defined, intransitivity in a choice task is likely a stronger indicator of lack of knowledge than in areas where a choice task relies on judgmental accuracy and choice transitivity (Funk et al., 2020).

Additionally, our proof result can be utilized to actively improve the efficiency of a choice protocol, particularly in prediction contexts where high accuracy is critical. By implementing either the *Random Search* or *Strategic Search*, one saves time and cost in eliciting expert responses, key factors when there is a small pool of expertise. As an alternative to ending a choice protocol early, active measurement of intransitivity in an expert's responses can be a point of discussion with expert respondents *before* the task is completed, saving time. It can be used to highlight if the alternatives in a cycle are indistinguishable due to the just noticeable difference in the expert's decision process, or if there is another source of uncertainty underlying the cyclic responses (Keeney & von Winterfeldt, 1989; Keeney, Raiffa, et al., 1993).

As graph size increases, the MFES becomes more difficult to calculate, as the problem is NP-hard. An open area of exploration that would improve this work would be to have upper and lower asymptotic bounds on the MFES as a function of graph size. There may also be a fruitful area of exploration looking at the path dependency of different structures of intransitivity. As a prediction task, and the resulting graph, grow larger, there are more possible paths to search the graph. There may be ways to use path information to improve the search algorithms proposed here.

Intransitivity is usually discussed in binary terms. Either intransitivity is an irrational behavior that demonstrates a flaw in thinking, or it is a natural part of the decision process and can be ignored. Our work demonstrates that intransitivity exists on a spectrum that can provide potentially useful information, mid-process and in the aggregate, about the knowledge and decision process of respondents.

# 4

# Modern preference learning with small data

## 4.1 Introduction

Tradeoffs between risks, costs, and benefits are part of any public decision (Fischhoff, 2015), whether it is the risk of patient death from an FDA approved obesity surgery treatment (M. P. Ho et al., 2015), or increase in the price of electricity to reduce harmful emissions (Sergi, Davis, & Azevedo, 2018; Sergi, Azevedo, Xia, Davis, & Xu, 2019). Members of the public must balance those tradeoffs when deciding to support (or oppose) public policies. However, there are significant barriers to accessing the relevant risk-cost-benefit information, expressing coherent preferences over alternatives given that information, and coming to a group decision that respects the many perspectives of others. In this paper we evaluate the capability of a novel neural network architecture to learn policy-relevant choice rules found in the behavioral decision sciences literature, comparing the architecture's performance against benchmark statistical and machine learning alternatives. Using large-sample simulations, we test the capability of these preference learning models to represent four types of behavioral choice rules (two variants of strong utility, additive difference, and multi-attribute linear ballistic accumulator). Using small-sample simulations, we consider not only the capability of representing behavioral choice rules, but the ability of learning those rules under the small data constraints imposed by human subjects data collections tasks.

This work is critical for public policy analysis because prior research has found that there is heterogeneity in preferences for public policies (De La Maza, Davis, Gonzalez, & Azevedo, 2018; Sergi et al., 2018, 2019). In the US, De la Maza, Davis, Gonzalez, and Azevedo find that clusters of decision-makers have opposed preferences in their willingness-to-pay for $CO_2$ abatement (De La Maza et al., 2018; Sergi et al., 2018), where a graph-based clustering method found that almost 30% of the sample had intransitive cycles in their choices, and many were unwilling to make trade-offs (De La Maza et al., 2018). In Chile, a pilot discrete choice survey by De la Maza, Davis, and Azevedo revealed 16 subgroups with heterogeneous preferences for environmental protection. Preference heterogeneity is also important at the technical frontier, where viewpoints differ among those who set priorities for research and development investments (Funk et al., 2020). To address this potential heterogeneity, we use methods that can learn choice rules without prior knowledge of their functional form. This builds on prior active preference learning work that uses adaptive experimental design methods for model testing when hypotheses are pre-specified (Cavagnaro, Myung, Pitt, & Kujala, 2010; Cavagnaro, Gonzalez, Myung, & Pitt, 2013; Kim, Pitt, Lu, & Myung, 2017).

### 4.1.1 Behavioral choice rules

Research in the decision sciences has enumerated threats to well-reasoned preferences that stem from basic psychological and cognitive processes (Busemeyer & Townsend, 1993; Bhatia, 2013; J. S. Trueblood, Brown, & Heathcote, 2014), particularly when preferences are expressed in complex and unfamiliar domains (Fischhoff, 2005). While there are many threats to the expression of coherent preferences, and many theories that explain those threats (Erev, Ert, Plonsky, Cohen, & Cohen, 2017), no theory is able to characterize choice under the many contexts facing the public, and decision-makers are often heterogeneous in both their values and the rules they use to make decisions (Fischhoff, 1991).

To allow for that heterogeneity, we focus on three important behavioral choice models for two-alternative forced choice tasks based on the empirical decision science literature, although extension to the multi-alternative case is often straightforward. In these tasks, the goal is to infer a choice rule from observations of forced choices between a left (L) alternative and a right (R) alternative given the attributes $x^L$ and $x^R$ of the left and right alternatives (respectively). The three types of behavioral choice rules are: 1) **strong utility** (Marschak, 1959; Block & Marschak, 1960; R. Luce & Suppes, 1965; McFadden et al., 1973; McFadden, 1999), 2) **additive difference** (Tversky, 1969), and 3) **multi-attribute linear ballistic accumulator** (MLBA) (J. Trueblood, Brown, & Heathcote, 2013; J. S. Trueblood et al., 2014). For the strong utility function (McFadden et al., 1973; McFadden, 1999), choice is the outcome of a Bernoulli random variable Bernoulli($P(L \succ R)$) where $P(L \succ R) = f(u(x^L) - u(x^R))$ is the probability of the left alternative (L) being chosen over the right (R), and $u(x^L)$ and $u(x^R)$ are utility functions over the attributes of the left ($x^L$) and right ($x^R$) alternatives in the choice set. If $f()$ is the logistic CDF, then $P(L \succ R) = \frac{e^{\alpha(u(x^L)-u(x^R))}}{1+e^{\alpha(u(x^L)-u(x^R))}}$. Second, the additive difference model, introduced by Tversky (1969) (Tversky, 1969), has real-valued functions $u_1, u_2, \ldots, u_K$ over attributes $k \in \{1, 2, \ldots, K\}$ and increasing continuous functions $\phi_1, \phi_2, \ldots \phi_K$ such that the choice of the left alternative over the right is $C(L \succ R) = \mathbb{I}\left[\sum_{k=1}^{K} \phi_k(u(x_k^L) - u(x_k^R))\right]$ where $\mathbb{I}(z) = 1$ if $z \geq 0$, and 0 otherwise. The lexicographic semiorder is one of the simplest and most well-studied additive difference models that allows intransitive preference (Tversky, 1969; Birnbaum, 2010; Davis-Stober, 2012; Bräuning & Hüllermeier, 2017). Third, context dependent preference models, such as the MLBA (J. Trueblood et al., 2013; J. S. Trueblood et al., 2014), allow the evaluation of alternatives to depend on the other alternatives in the choice set. In the MLBA, choice is determined by a drift-diffusion process crossing a threshold, where the mean drift rate $d_i$ for alternative $i$ is based on pairwise comparisons with the other alternatives. For example, in the three alternative case $d_1 = V_{12} + V_{13} + c$ where $V_{12}, V_{13}$ are comparison valences of alternative 1 to the two other alternatives. The general form of $V$ is $V_{ij} = \sum_{k=1}^{K} w_k(x_{ki}, x_{kj}) \times \left[u_k(x_{ki}) - u_k(x_{kj})\right]$ where $w_k(x_{ki}, x_{kj})$ is a weighting function for attribute $k$ and $u_k$ is a function that subjectively transforms attribute $x_k$. For example, Trueblood, Brown, and Heathcote (J. S. Trueblood et al., 2014) use the similarity function $w_k = e^{-\lambda_k |u_k(x_{ki}) - u_k(x_{kj})|}$ for weights and a power function $u_k(x_k) = (\frac{x_k}{a_k})^{m_k}$ for the subjective transformation of attributes.

### 4.1.2 Machine learning approaches

Parametric learning methods, such as the multinomial or mixed logit models, assume the functional form of individual preferences, then find the parameters that best fit choices given that functional form. This works well when the choice data are well characterized by a simple theoretically-motivated model, but can lead to inaccurate results when there is a mismatch between behavior and theory. More flexible learning methods, such as neural networks with hyperparameter optimization, are an alternative approach that offer flexibility beyond simple parametric models. However, neural networks often require more data, and do not have efficient closed form solutions, instead requiring approximation (Judd, 1990). This is important for human decision-making tasks, where data are limited by the number of respondents that are willing to participate, and the amount of time each respondent is willing to contribute. For example, to avoid respondent fatigue, tests of Prospect Theory use no more than a few hundred observations (Broomell & Bhatia, 2014). On the other hand, modern language models require millions or even billions of observations (Brown et al., 2020; Sun, Shrivastava, Singh, & Gupta, 2017).

In theory, many machine learning models can represent arbitrary continuous functions (Hornik, 1991), yet they vary significantly in their learning efficiency. Efficiency improvements can be gained with the use of active learning, that selects alternatives to maximize the information gained about the target functions (Gal, Islam, & Ghahramani, 2017). Some approaches attempt to learn relationships with little-to-no data in the target domain, for example few-shot learning (Vinyals, Blundell, Lillicrap, Wierstra, et al., 2016), and transfer learning (Weiss, Khoshgoftaar, & Wang, 2016).

Specifying the underlying principles, or invariants, of choice that hold across many key decision-making models can significantly reduce the space of decision rules learnable by the network, and thus the amount of data required to train the model (Wolpert & Macready, 1997). Here we focus on those behavioral invariants to constrain the neural network structure to 1) learn a variety of behavioral choice models and 2) do so with the limited data sets realistic to choice tasks. Consider the invariants across strong utility, additive difference, and MLBA. One is *subjective transformation of single attributes*, that single attributes are transformed psychologically when considered in choice. For example, in MLBA the function $u_k(x_k)$ transforms the observed attribute $x_k$ through some function $u_k$. Second, is *single attribute weighting*. In the strong utility function, those weights are additive and independent of the attribute levels. In the MLBA and additive difference models, the weights depend on the attributes themselves ($w$ in MLBA and $\phi$ in additive difference). A third invariant is *pairwise comparison processing of alternatives*, so the model should be decomposable based on pairwise comparisons. For example, in MLBA, the functions are summed over pairs of alternatives $d_i = \sum_{j \neq i}^{n} V_{ij}$. While a densely connected network can learn any type of choice rule (with enough units) (Goodfellow, Bengio, Courville, & Bengio, 2016), dense networks do not take advantage of these invariants. To improve upon the densely connected network, we propose a *twinned neural network* (Chopra, Hadsell, & LeCun, 2005; Rao, Wang, & Cottrell, 2016; Bromley et al., 1993) architecture that learns an identical encoding of each alternative (Behler & Parrinello, 2007; Behler, 2011) (see Figure 4.1). The first $N$ layers of the network *subjectively transform single attributes* of an alternative using non-linear activations, with shared weights $(W_1, \ldots, W_n)$ across all alternatives. Single attributes are then differenced across alternatives (*difference layer*), weighted (*attention weight layer*), then a non-linearity is applied and multiplied by the difference (*single attribute value layer*). Summation across all attributes and application of the logistic function then determines choice (*total value layer*). By varying the activation functions and dense connections after the first $N$ layers, this architecture can likely represent many types of behavioral choice rules, including the strong utility, additive difference, and MLBA.

We test our proposed neural network architecture in a variety of synthetic experiments as well as real contexts from prior research. First, we train and tune the architecture on synthetic data sets much larger than traditional choice tasks in order to understand its theoretical limits. Second, we test the new approach's ability to learn on small synthetic datasets. Finally, we use data and models from two choice experiments to test the approach in real environments at the technological frontier.

## 4.2 Methods

Here we describe the neural network architecture appropriate for modeling two-alternative discrete choices. In order to examine the capabilities of the network structure (Figure 4.1), we first test it on a large, simulated data set generated from a variety of prominent behavioral choice functions. The network is evaluated against traditional methods (parametric and non-parametric) for out-of-sample prediction as well as flexibility across applications. We then test the model's capabilities on small, simulated data sets that reflect the size of realistic choice tasks.

### 4.2.1 Network structure

The network structure leverages the aforementioned invariants in the two-alternative forced choice pairwise comparison task. For this task, a learning model should be invariant to the presentation order of the alternatives (i.e. the learned function $f$ should not be substantively different if we evaluate $f(X^L, X^R)$ or $f(X^R, X^L)$). To create this property in the network we use a twinned structure (Bromley et al., 1993). This forces the weights from both inputs to be identical and train in tandem. A dense layer (or layers) is added to accommodate any non-linearities between objective and subjective attributes of the alternatives (as seen in both the MLBA and additive difference models) (J. S. Trueblood et al., 2014; Tversky, 1969). A differencing layer creates the comparison between inputs. This is followed by another dense layer section for non-linear transformations of the difference, as well as the option for inclusion of the output from the difference post-transformation (as in the MLBA case). Lastly we compute a total probability output corresponding to the resulting choices.

### 4.2.2 Behavioral data generating processes

We generate data from four common behavioral models of choice: MLBA (J. S. Trueblood et al., 2014), linear, additive difference (Tversky, 1969), and ideal point (Coombs, 1964). These models represent a range of complexity in the functional forms and the choice outcomes possible for a discrete choice task. MLBA is the most flexible of the decision models capturing context effects well in multi-attribute, multi-criteria choice including time deliberation effects (J. S. Trueblood et al., 2014). The linear model has a simpler, closed computational form and fits many normative theories of rational choice (Tversky, 1969). The additive difference model is broader than the linear model as it reduces to the linear model when the transformation function is linear, but it allows for a broader set of transformation functions that can induce intransitivity (Tversky, 1969). We examine two cases of the additive difference model. The first utilizes a hyperbolic tangent transformation of the differences to understand the implications of nonlinearities. The second is the lexicographic semiorder, a model that is well-known for inducing intransitivities (Tversky, 1969). Going forward we refer to the hyperbolic tangent model as the additive difference model and use the term

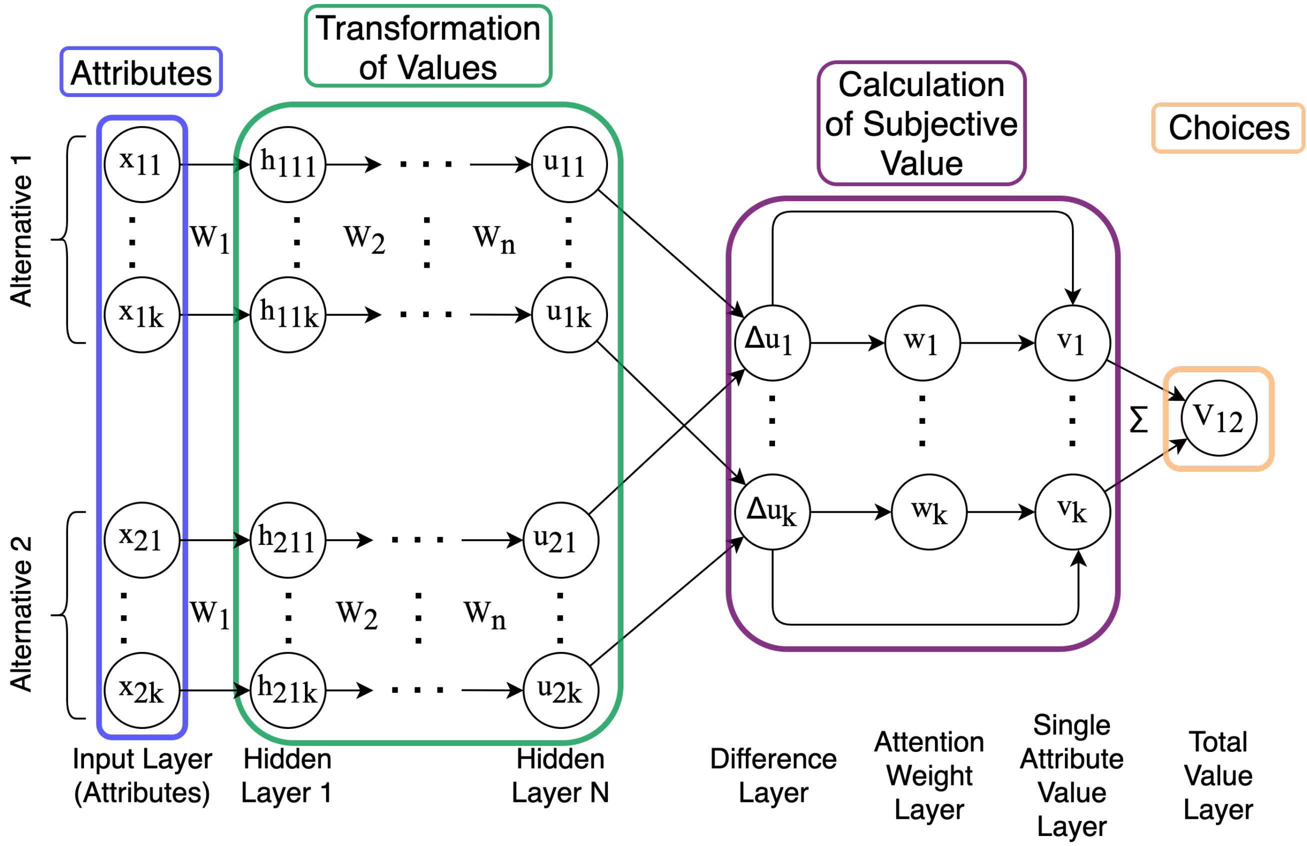


Figure 4.1: The architecture for the network incorporates elements to enforce theoretical constraints of the choice models.

lexicographic semiorder for this specific case. Lastly, the ideal point model (Coombs, 1964) represents a common case where there is an ideal level of each attribute, and deviation from that ideal point is penalized in the valuation of the alternative.

### 4.2.3 Model formulations

**MLBA**

For generating the MLBA data, we begin with a linear transformation of objective ($X$) to subjective ($u$) weights.

$$u(X) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K$$

Valences for a pairwise comparison of left and right options ($L$ and $R$) follow a weighted sum of differences equation.

$$V^{LR} = w_1^{LR} * (u_1^L - u_1^R) + w_2^{LR} * (u_2^L - u_2^R) + ... + w_k^{LR} * (u_k^L - u_k^R)$$

Weights for the differences use a $\lambda = 0.1$ times the absolute value of the subjective attribute difference.

$$w_i^{LR} = exp(-\lambda |u_i^L - u_i^R|)$$

The logistic function of the valence (below) is then used as the probability in a binomial generator to produce the binary choice data for training.

$$P(y_t = 1 | X^L, X^R) = \frac{e^{V^{LR}}}{1 + e^{V^{LR}}}$$

### Linear

For the linear data generating process, we use a linear model of the attribute differences.

$$u(X^L - X^R) = \beta_1(x_1^L - x_1^R) + \beta_2(x_2^L - x_2^R) + \cdots + \beta_k(x_k^L - x_k^R)$$

This difference is the input for a logit transform to generate probabilities for a binomial data generating function to create the training data.

$$P(y_t = 1|X^L, X^R) = \frac{e^{(u(X^L - X^R))}}{1 + e^{(u(X^L - X^R))}}$$

### Additive difference

For the additive difference data generating process we use a hyperbolic tangent transformation of the objective attributes rather than a linear.

$$u(X^L - X^R) = \sum_{k=1}^{k}(\tanh[x_k^L - x_k^R])$$

These subjective differences are then used in the logit transform to produce probabilities for binomial data generation as in previous steps.

$$P(y_t = 1|X^L, X^R) = \frac{e^{(u(X^L - X^R))}}{1 + e^{(u(X^L - X^R))}}$$

### Lexicographic semiorder

For the lexicographic semiorder we use a special case of the additive difference model where

$$\phi_k = I\left[u(x_k^L) - u(x_k^R) \geq \epsilon_k\right]$$

which is an indicator function that is 1 if $u(x_k^L) - u(x_k^R) \geq \epsilon_k$ and 0 otherwise. Here $\epsilon$ is called the *just noticeable difference* (JND) to indicate that the decision-maker ignores the difference between L and R on attribute $k$ unless that difference is sufficiently large (greater than $\epsilon_k$). For example, two cars that differ by \$10 or even \$100 on their price would be treated as having roughly equal price (i.e., \$100 $< \epsilon_k$). In the lexicographic semiorder a decision-maker sorts the attributes of alternatives from most to least important, then chooses the alternative that is best on the most important attribute unless that attribute is within a just-noticeable difference of the second-best alternative. This process is repeated on the next most important attribute until a decision is made.

### Ideal point

For the ideal point data generating function there is an ideal level for each attribute. The objective attributes are transformed by the square of the difference of each attribute from this ideal level $d_k$.

$$u(X^L - X^R) = \beta_1 \left[(x_1^L - d_1)^2 - (x_1^R - d_1)^2\right] + \cdots + \beta_K \left[(x_K^L - d_K)^2 - (x_K^R - d_K)^2\right]$$

Probabilities for generating binomial data are calculated as from the previous generating processes.

$$P(y_t = 1 | X^L, X^R) = \frac{e^{(u(X^L) - u(X^R))}}{1 + e^{(u(X^L) - u(X^R))}}$$

### Training the twinned network

All code is written with Keras and Tensorflow in Python (Chollet et al., 2015). The network parameters are tuned on the MLBA data. We generate 10,000 3-dimensional MLBA pairwise comparisons and train the model on a 40/40/20 train/validate/test split of this set. Model parameters are tuned using the Talos package ("Autonomio Talos [Computer software]", 2019) with adjustments in unit width, layer depth, regularization levels, and training epochs. Once the parameterization of the model was selected, the same parameters and hyperparameters are used for each instance of the twinned network.

## 4.3 Experiment 1: Large data

Initially we test the model on a large data set of 10,000 pairwise comparisons (the same volume as for the tuning process) to see if the twinned network is able to flexibly learn across behavioral model contexts. The best model fit for the twinned network on the MLBA pairwise comparison data has an area under the receiver operator characteristic curve (AUC) of 0.96. The network learns the data generation process of the MLBA function with the large data set. We then test the same parameterization of the twinned network on the linear, additive difference, lexicographic semiorder, and ideal point model data. AUC for these data generating processes is 0.92, 0.77, 0.68, and 0.93, respectively. The twinned network is flexible enough to learn two of the four data generating processes, but does not perform as well on the additive difference or lexicographic models.

One explanation for the differences in learning is the relative smoothness of the behavioral models. The linear and ideal point models change linearly and quadratically across the entire range of differences, while the hyperbolic tangent function saturates at +1 and -1, making the dynamic range of the function of the differences narrower than the previous two models. The lexicographic semiorder is even more extreme in this respect, as it is an indicator function at the just-noticeable difference. Given the twinned network relies on backpropagation for training, these “less smooth” functions may require more data, or data concentrated in different training regions, to improve the network’s fit.

## 4.4 Experiment 2: Small data

For the small data experiment, we aim to understand the twinned network’s performance on data sizes that were scaled to those plausible for single respondent or small, multi-respondent aggregated choice surveys. We measure the performance on data sets of 50 to 500 pairwise comparisons, generating data using the methodology outlined in *Model Formulations*. For each data set we repeat the random 40/40/20 division of training/validation/testing data 50 times to see how the volatility of the learning process is affected by varying inputs.

Figure 4.2 shows the results from these training runs. For the smallest data sets tested, similar

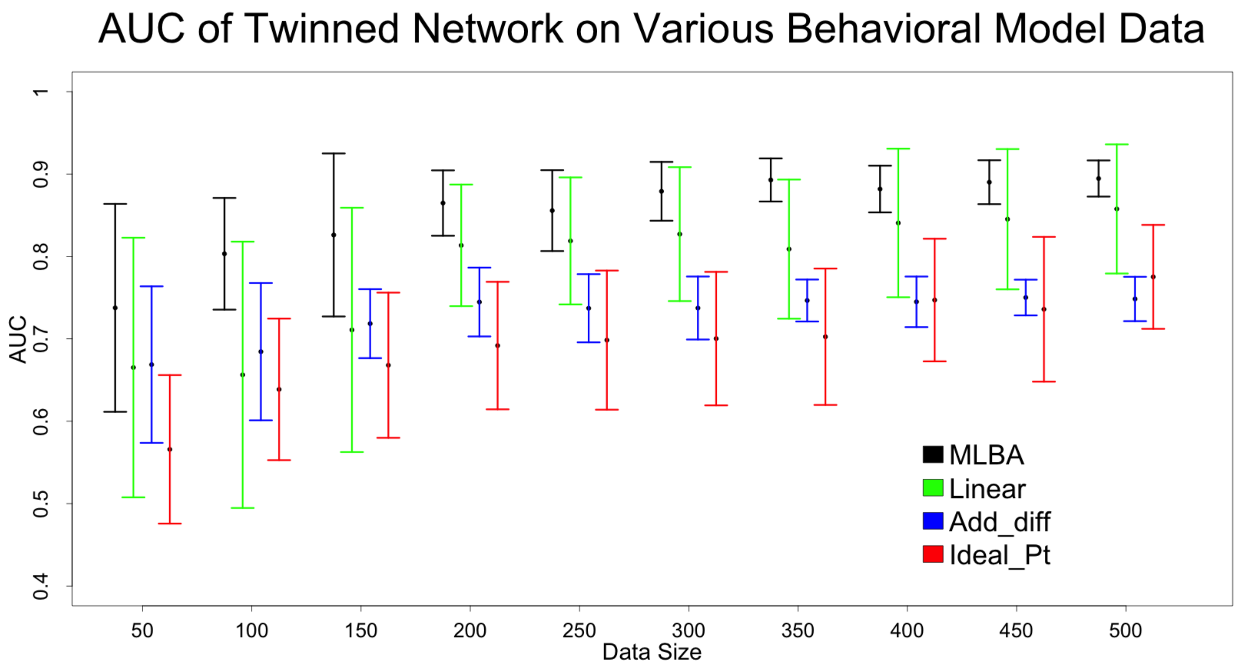


Figure 4.2: Simulated data was generated using the MLBA, linear, additive difference, lexicographic semiorder, and ideal point behavioral models. We use 50 repetitions of training the models with random partitions of data between training, validation, and testing to highlight the range of learning possible. Error bars represent one standard deviation from the mean.

to sizes of a single respondent in a choice task, the twinned network has an average AUC near 0.7 for all data generating processes and a high degree of variance (including 0.5 within one standard deviation from the mean for the additive difference and lexicographic models, making it no better than chance). As the sample sizes approach those closer to a repeated choice task or an aggregated choice task amongst a group of respondents (200 – 500), the twinned network averages an AUC above 0.85 for the linear and ideal point models, but the MLBA, additive difference and lexicographic models all remain at or below 0.8. There is a noted improvement in learning the MLBA as you move from these training sets to the large data in Experiment 1, while it appears there is little improvement in the additive difference and lexicographic models even with significant training data increases. The ideal training data set size is likely between Experiment 1 and Experiment 2 for the linear and ideal point models, while the ideal data set size for the additive difference and lexicographic semiorder may be much larger, and of a different composition.

## 4.5 Experiment 3: Method comparison

In order to position the twinned network model in the landscape of potential methods a researcher could use, we compare it to traditional analytical methods, both parametric and non-parametric. We compare multiple methods' peak performance as well as the spectrum of performances across the suite of behavioral models. We train the twinned network, a standard logistic regression model, a random forest (T. K. Ho, 1995), and gradient boosting (Friedman, 2001) on the four data generating functions. Code for the regression, random forest, and gradient boosting methods use python package scikit-learn (Pedregosa et al., 2011). In this way we compare the overall flexibility of our approach.

Capturing a wide range of data scenarios is critical to verifying the usefulness of our new method. We vary the complexity of the data generating processes across the five behavioral models to capture this effect. The range of possible objective attribute values in the sampling space is one way to do this. When the attribute range is small, the alternatives have smaller differences and the resulting sampling probabilities are more uniform (near 0.5). This leads to noisier data that is more difficult for model training. With larger attribute ranges, the alternative differences are also more likely to be large and lead to bimodal probability distributions with peaks near 0 and 1. For the additive difference model with the hyperbolic tangent non-linearity, however, the wider data attribute range *does not* create strictly less noisy data, but rather concentrates probabilities near 0, 0.25, 0.75, and 1 (with the 3 dimensional data used in our study). Due to the different transformations from objective attributes to final choice probabilities with each behavioral model, we chose the attribute ranges to qualitatively capture similar distributions of prediction probabilities. The x-axis in Figure 4.3 has been condensed represent this categorical format.

Figure 4.3 shows the results of the four statistical models trained on four of the behavioral models for varying data complexity. Here we use a large sample size ($N = 10,000$) to isolate the data complexity from the data size constraints of Experiment 2. The MBLA and linear graphs behave as we predict, with increasing AUC as data attribute ranges move from small to large (decreasing in noise). For the additive difference model with a hyperbolic tangent non-linearity, the twinned network and logistic regression perform the best when the comparisons were relatively uniform in probability. As the distribution of probabilities becomes more multi-modal, the twinned network and logistic suffer and the random forest model improved significantly. The hyperbolic tangent function saturates at -1, and 1 with large attribute levels. A decision space analogy is that each attribute difference becomes either a pro (+1) or a con (-1) and those pro and con categories are then summed, where the alternative with the most "pro" counts is the final choice. It is possible that the random forest ensemble captures this multidimensional space of dividing hyperplanes better than the logistic or twinned network. The ideal point process shows all models doing similarly when the data is uniformly noisy, but the twinned network did much better than the other models when the data attribute range increases.

Figure 4.4 shows the results from the data difficulty experiment with the lexicographic semiorder. In this case, the method used for increasing the difficulty in fitting the model is to increase effective noise through the JND. As the JND increases, the difference between attribute levels needs to be larger for that attribute to force the choice between alternatives. In our simulation using integer values for the attribute levels, a JND of one means any difference between attribute levels creates a choice. This means the model will always select the alternative based on the first attribute, unless they are identical. As we see in Figure 4.4, this creates a behavioral model that is easy to learn as it very closely resembles a single attribute linear model. As the JND increases, the difference for each successive attribute must be larger to force a choice. At a JND of 5 (in our simulation), the JND is larger than any attribute level differences and the choices are entirely random. In between we see that all the models steadily decrease with the JND change. The logistic regression may be slightly worse than the other models, but we caution that this difference may be too small to warrant concern given the many benefits of using a well-developed modeling tool.

## 4.6 Experiment 4: Practical application

### 4.6.1 Sergi et al. (2018) data

In order to test our new framework in an empirical setting, we use data from Sergi *et al.* (2018) (Sergi et al., 2018). This study looks at individuals' preferences for alternative energy production portfolios based on climate and health information. Respondents are presented pairwise

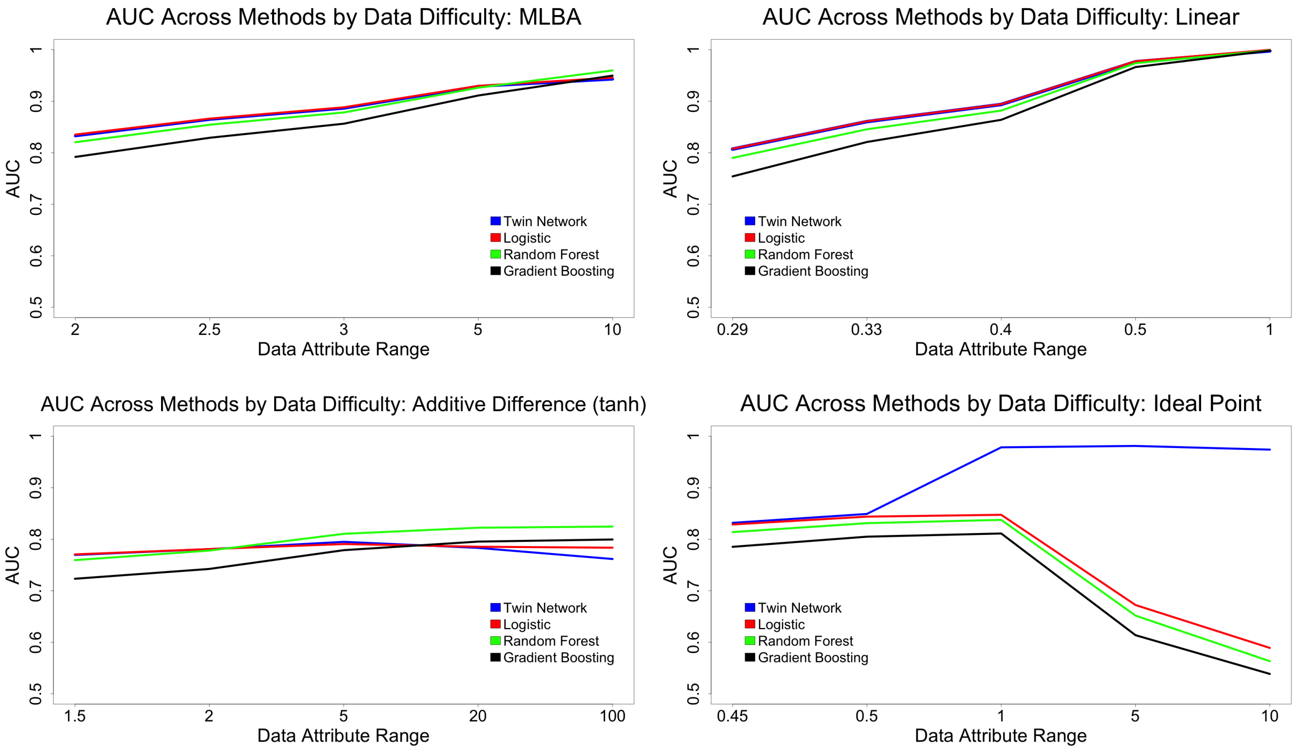


Figure 4.3: As the attribute space for simulated alternatives widens, more extreme differences exist between pairs. This creates data sets that are more noisy when the range is small (left of graphs) and less noisy when the range is large (right of graphs) as generated probabilities are closer to 0 and 1.

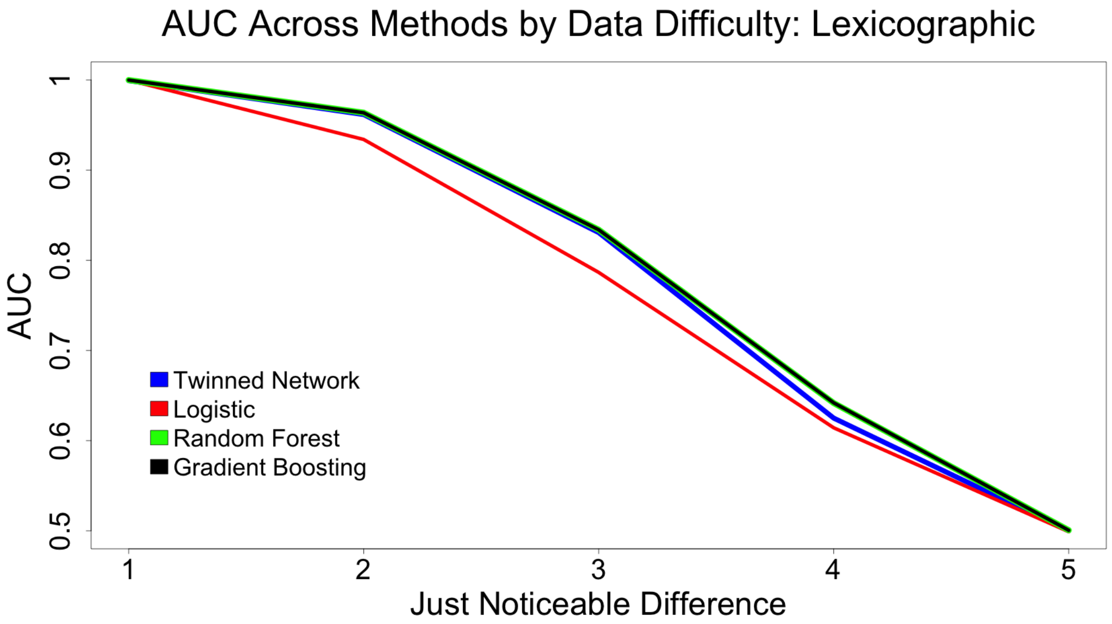


Figure 4.4: For the lexicographic semiorder, noise in the simulation is increased by increasing the just noticeable difference (JND). As the name implies, a larger JND means the attribute differences have to be larger to register a preference. Otherwise, that attribute is skipped in the decision process, and with a large enough JND, the choice is always random.

comparisons between two energy portfolios varying the dominance of particular production technologies (coal, nuclear, natural gas, renewables, and efficiencies), monthly energy bill, CO2 pollution and SO2 pollution.

We train the twinned network on the 3, 536 pairwise comparisons within the Sergi data set. Data is randomly split as in the simulations in a 40/40/20 training/validation/testing regime. We also fit a logistic regression model as a baseline for comparison with the same testing data withheld. With 30 repetitions, the average AUC is 0.91 for both the twinned network and the logistic regression. The twinned network does as well as, but not better than, the logistic regression model at learning preferences and predicting choices in this real-world data set. This aligns with the previous simulation results.

### 4.6.2 Funk et al. (2020) simulation

Next we test the twinned network on a data set where we have more information about the models of the individual decision makers. We use the experts from Funk *et al.* (2020) (Funk et al., 2020) as

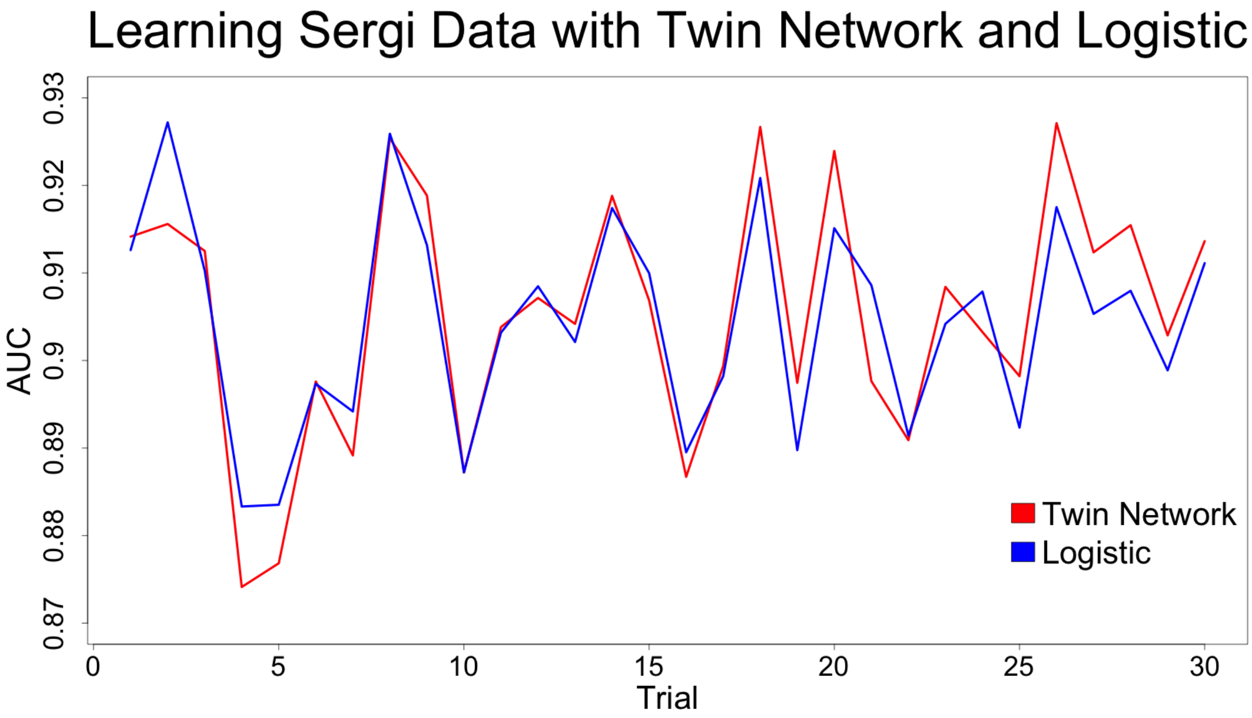


Figure 4.5: The twinned network predicts the holdout test set as well as *but not better than* a traditional logistic model for the Sergi *et al.* data. Minor variation between the logistic and twinned network do not indicate the new approach is a consistent improvement.

the model for simulating a heterogenous set of decision makers. In Funk *et al.* (2020), experts in additive manufacturing for aerospace were asked questions about the comparative feasibility between pairs of a set of jet engine parts. Their part preferences were then modeled in aggregate and at the individual level. Experts used a variety of behavioral models to make their choices in the study. Of those experts, 13 of 27 were well characterized by a linear model of the attributes, three used a single attribute decision model, and ten used decision trees to make their choices. We generate data to simulate this aggregate group to test our network on the aggregate as well as the various individual models from the study.

We generate data from 27 experts using the models characterized in the paper (linear, single attribute, and the three decision trees detailed in the SI). For the aggregate training we train the model on the total response set from 27 experts with 66 pairwise comparison responses each. We then test the twinned network and a logistic regression model on this aggregate set. The twinned network test AUC is 0.63 and the logistic model AUC is 0.68. These poor learning rates are not surprising given the diversity of behavioral models within set.

Next we train each model on one expert's simulated 66 pairwise comparisons, and test them against a set of 66 pairwise comparisons generated from their own behavioral model. We use the same statistical methods from the simulation experiments (twinned network, logistic regression, random forest, and gradient boosting) on each behavioral model type from the expert set. Since the decision tree experts exhibited quasi-complete separation in their responses, traditional logistic regression would not be possible as their choices were deterministic. Here we use regularized ridge logistic regression as a base comparison.

Figure 4.6 shows the results of this study. All models perform well on the single attribute expert models. The regularized regression outperforms the twinned network on all categories (similar to the results from Experiment 2 given the similar data sizes). The random forest and gradient boosting perform poorly on the linear model, but do well on the decision trees as one might predict given they are based on methods to divide hyperplanes in a tree-like fashion. In general the logistic regression performs the best across the balance of the expert models while the twinned network does not appear to provide any improvements. Given the twinned network was not an improvement over logistic regression in Experiment 2 with similar data sizes, this result follows the

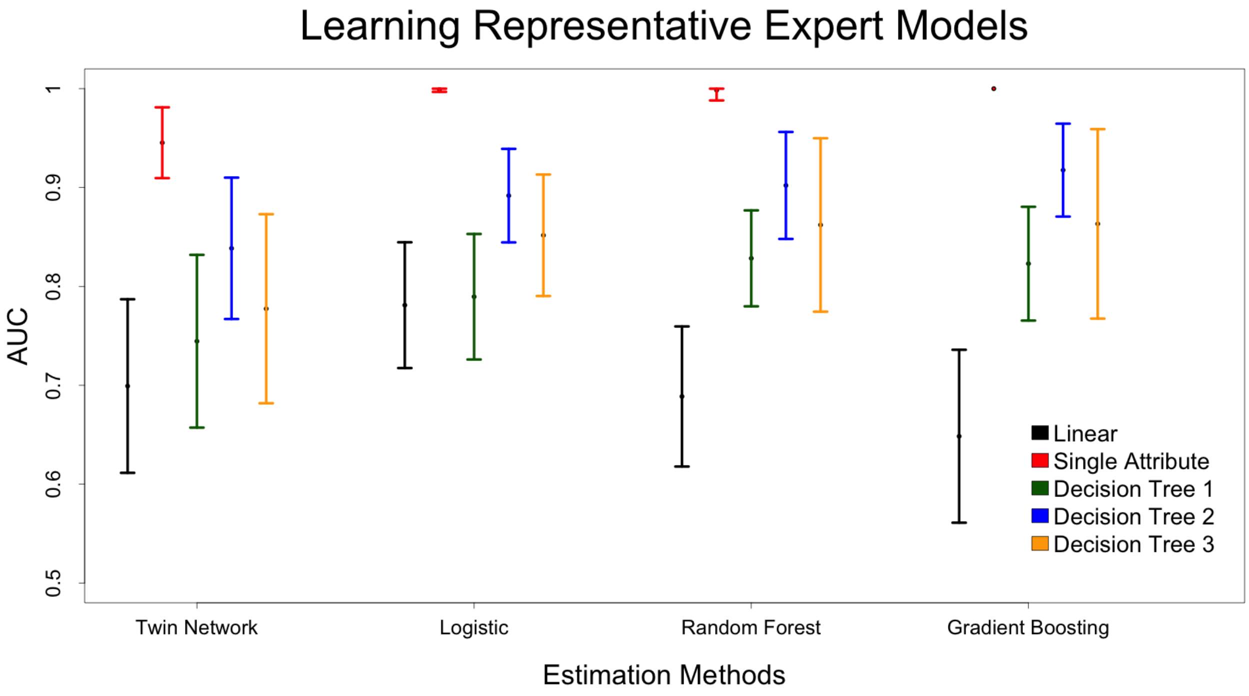


Figure 4.6: The twinned network is compared to traditional parametric and non-parametric estimation methods on data generated from Funk *et al.* 2020. The random forest and gradient boosting perform well on the decision tree and single attribute data, but the worst of the methods on the linear model. Logistic regression does relatively well across all models. The twinned network performs poorly learning these models in comparison to the traditional methods.

same reasoning. Likely the data range required for the twinned network to perform well is closer to the Sergi data in the thousands than the Funk data in the tens.

### 4.6.3 Future work: active learning

Given the questions around the effect of training data seen in Figure 4.2 and discussed in Section *Small Data*, as well as the results from the empirical test in *Experiment 4*, it would be informative to test an active learning framework for the twinned network. Active learning would utilize the second to last layer method by Geifman and El-Yaniv (Geifman & El-Yaniv, 2017) to inform the selection of new alternative pairs. We expect that this will lower the data requirements for model fit across data generating processes, but it may also lead to improved model performance.

## 4.7 Discussion

Heterogeneity across preferences exists in a variety fields that affect public policy and public well-being (M. P. Ho et al., 2015; De La Maza et al., 2018; Sergi et al., 2018, 2019; Funk et al., 2020). Measuring this heterogeneity and learning the degree to which it should be incorporated into public policy decisions requires flexible analytical techniques. Many of the standard parametric analytical methods suffer from limitations when there is a mismatch between decision makers' internal choice process and the statistical method chosen by a researcher. Our twinned network model provides a model-agnostic approach to measuring heterogeneous preferences.

Our twinned network's structure is informed by the theoretical elements that are shared across choice paradigms. The twinned first layers for left-right invariance and potential objective-to-subjective attribute transformations, the difference layer for comparing alternatives, and a final total value layer to accumulate the attributes into a choice probability are all consistent across the MLBA, linear, additive difference, and ideal point models. Additional layers for calculating subjective value output allow for the network to be complex enough to fit MLBA data, without *requiring* the non-linear elements from that process to be present.

We demonstrate the network's theoretical potential on large data generated from each of the four behavioral choice models, confirming that it is capable of learning all but the additive difference model well. On smaller data sets (as seen in Figure 4.2), the network did not perform as well. The twinned network reached only $AUC = 0.9$ for the linear and ideal point models, while the additive difference, lexicographic, and MLBA models fell well short. Additional work is warranted here to understand if this is a fundamental limitation of training a flexible model of this type, or additional tuning is possible to increase the small data learning capacity for out network.

As data complexity increases, the twinned network performs well relative to the other parametric and non-parametric methods we tested. All models perform relatively well across the noise levels tested for the MLBA and linear data generation. For the additive difference model, none of the methods reach an AUC above 0.85, though the random forest model does improve slightly across the range examined, while the other twinned network and logistic regression get worse. With the ideal point data, the twinned network matches the best method with high data noise and drastically outperforms the other three methods with low noise.

Lastly we test the twinned network on the data by Sergi *et al.* 2018 (Sergi et al., 2018) and Funk *et al.* 2020 (Funk et al., 2020). Here the model performs similarly to logistic regression but does not appear to offer marked improvements. The model is flexible and could provide benefits for active learning of choice data, but further refinements will be required to warrant its use over traditional methods.

# 5

# Conclusion

The technical frontier presents many challenges for experts, the industries they exist within, and the larger economies that rely on the progress and innovation of those industries. Chief among these challenges is uncovering the most productive directions for experimentation and innovation. The cues experts use to sort through the many potential paths are a mix of both tacit and explicit, which are often difficult to teach or evaluate in the moment. Codifying experts' knowledge through careful examination of simulated choice environments may be one way to improve the pace of innovation at the technical frontier.

This thesis presents one path towards this codification process. In a case study, theoretical proof, and computational experiment, I present a framework and tools for understanding expert choice at the technical frontier. In this setting outcomes for validation are difficult to obtain and there exists a large degree of heterogeneity in the structure and process of choices of experts. Despite these challenges, I demonstrate in the context of metal additive manufacturing for aerospace that knowledge can be gathered from this diverse expertise and productively applied more generally to the technology.

Beyond expert choice at the technical frontier, this framework can be applied for modeling many choices where tacit and explicit factors blend. Uncertainty about future outcomes or heterogeneity of choice model are factors that are not isolated to the technical frontier, but affect many areas of public policy and our daily lives (Sergi et al., 2018, 2019; Fischhoff, 1991; De La Maza et al., 2018). From patient preference to political forecasting, these tools improve our ability to capture the knowledge and preferences of individuals in contexts of uncertainty and risk.

# 6

# SI for Individual inconsistency and aggregate rationality

## 6.1 Supplementary figures

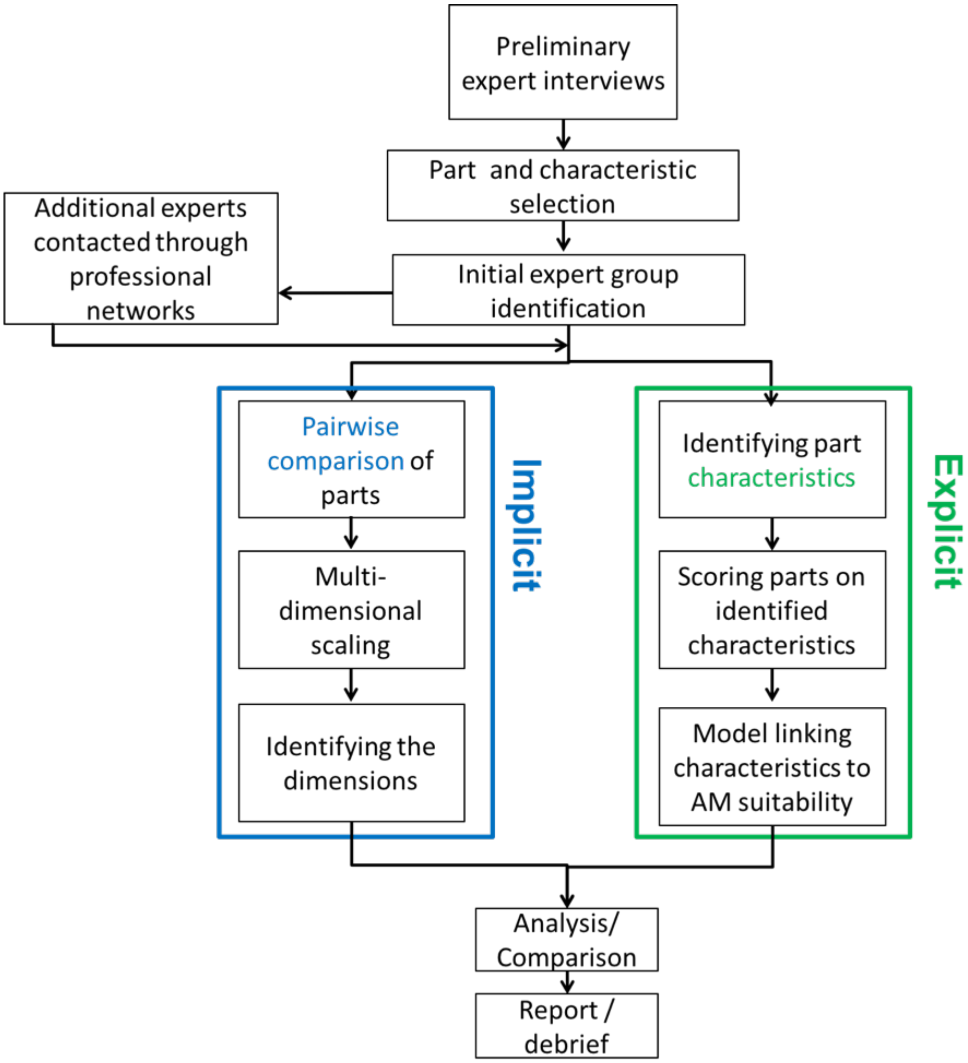

Preliminary expert interviews
Part and characteristic selection
Additional experts contacted through professional networks
Initial expert group identification
Implicit
Pairwise comparison of parts
Multi-dimensional scaling
Identifying the dimensions
Explicit
Identifying part characteristics
Scoring parts on identified characteristics
Model linking characteristics to AM suitability
Analysis/ Comparison
Report / debrief

Figure 6.1: A flow diagram depicts the initial sample collection, two section survey method, and data analysis of the research project.

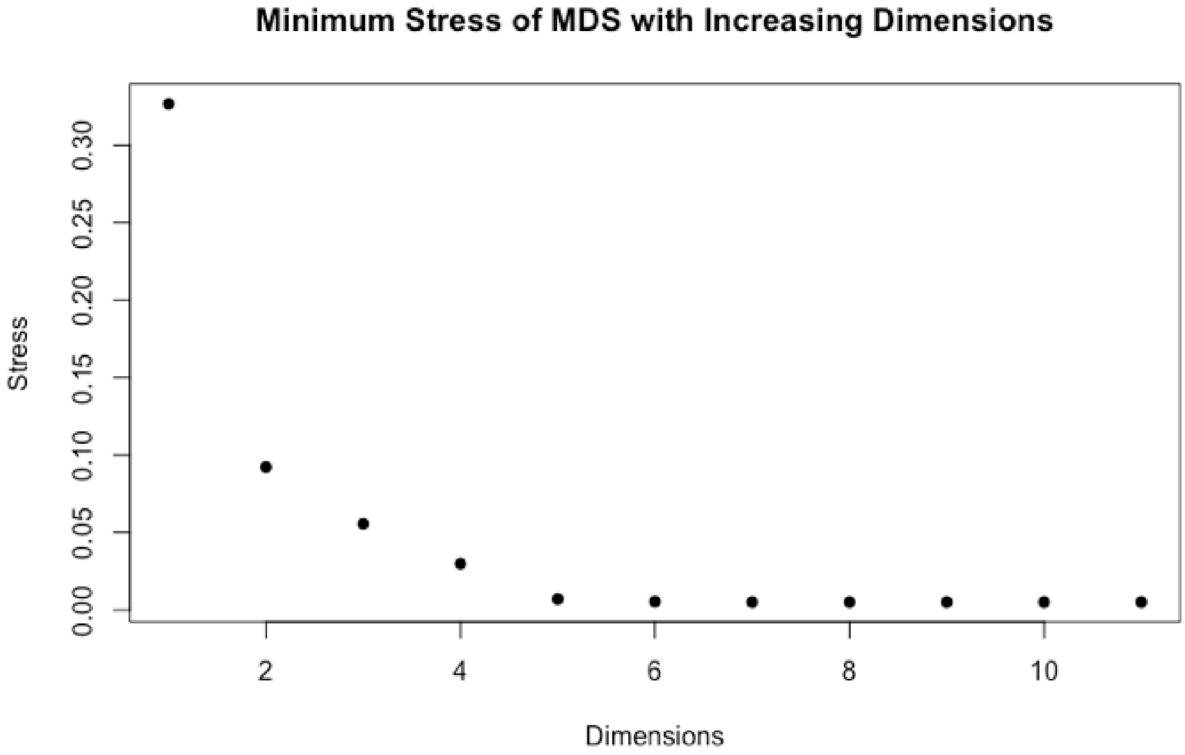


Figure 6.2: Successively running the MDS algorithm with varying dimensions shows the decrease in stress with added dimensions.

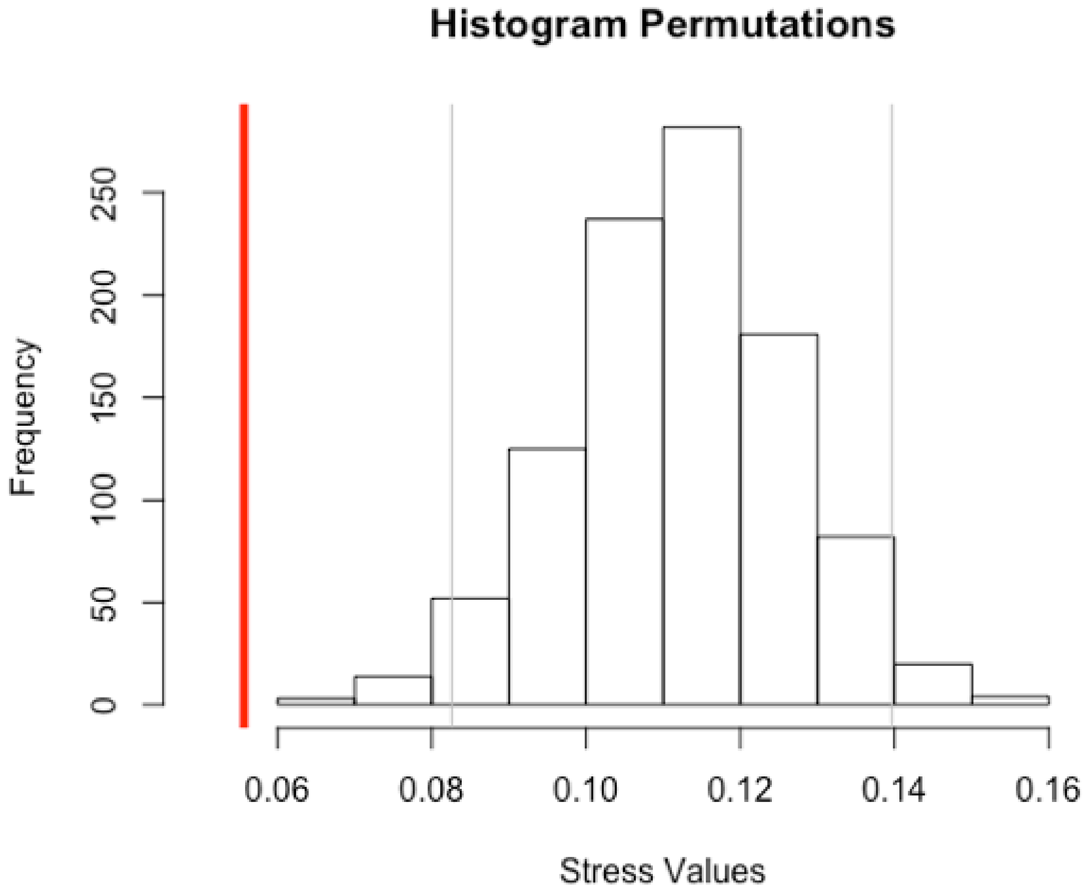


Figure 6.3: The stress value (0.056) of the data falls far outside the distribution (95% bounds given) of stress values found from permutation.

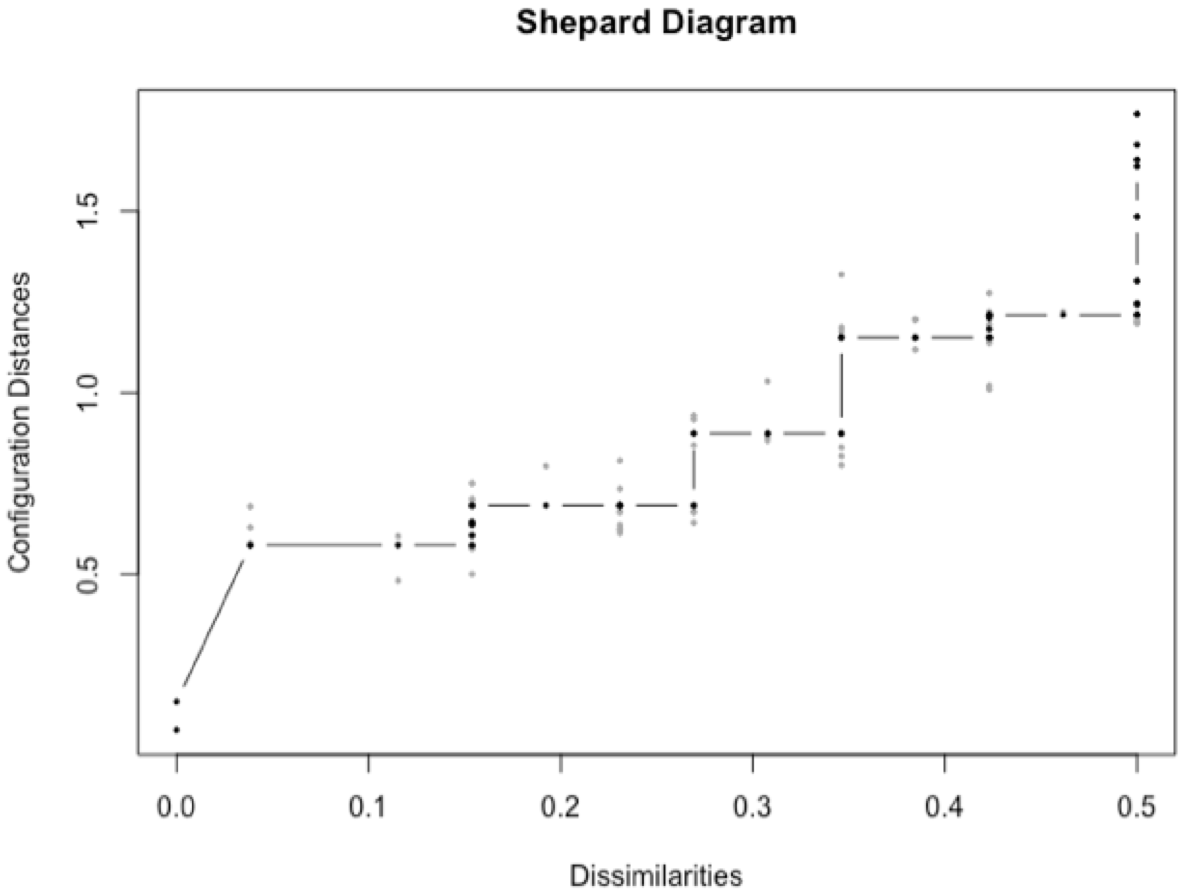


Figure 6.4: Shepard plot of 3D MDS result

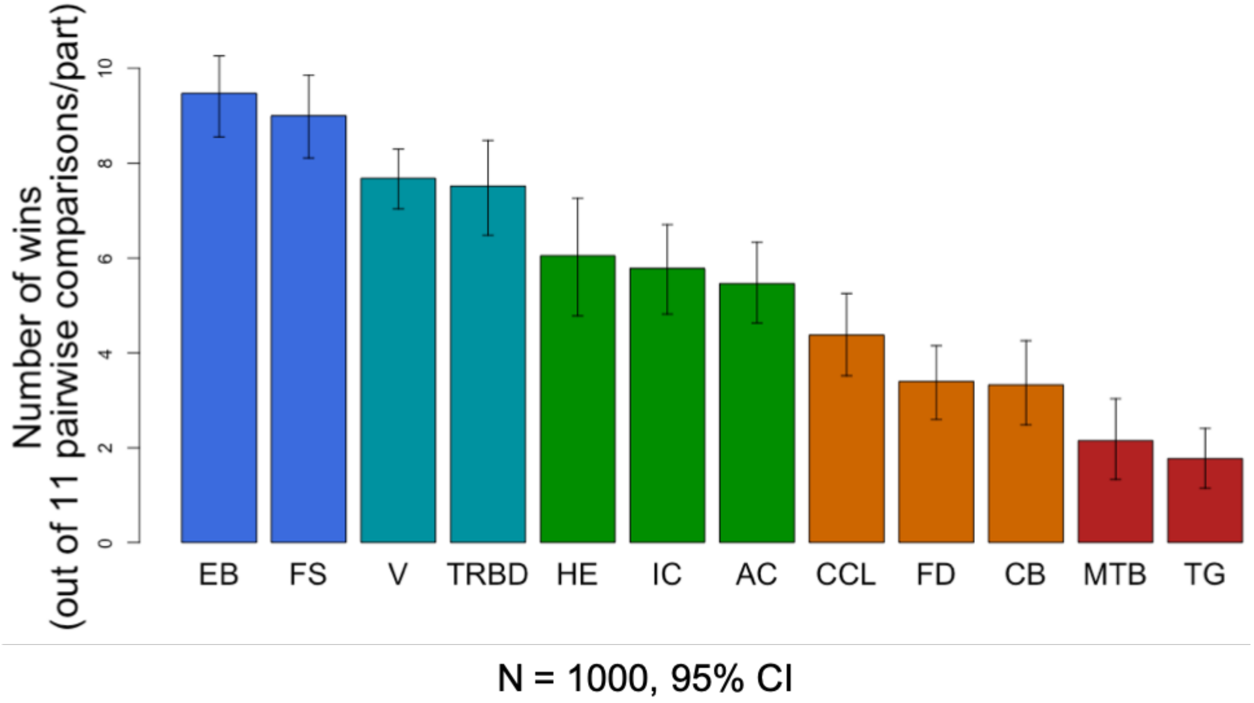


Figure 6.5: Graph of statistically significant groupings of the aggregate preference order

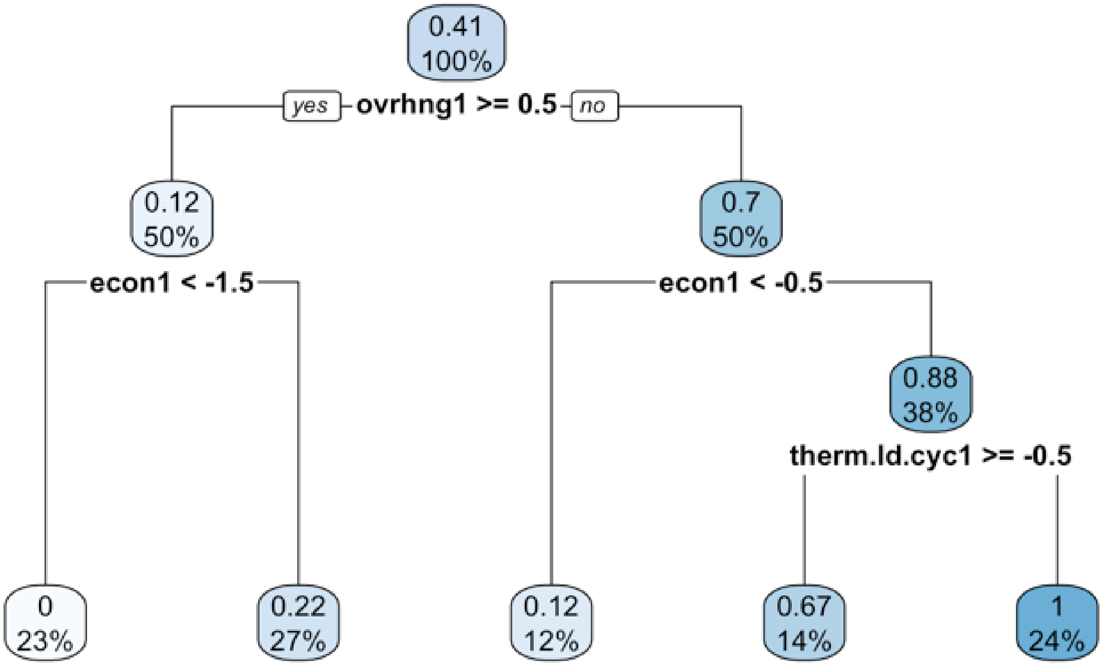


Figure 6.6: Decision tree for an expert using three attributes. Percentages indicate what percent of the overall choices are in a given branch of the tree while decimals represent the expected value of a given branch (where 1 represents choosing the part on the left and 0 choosing the part on the right in a pairwise comparison)

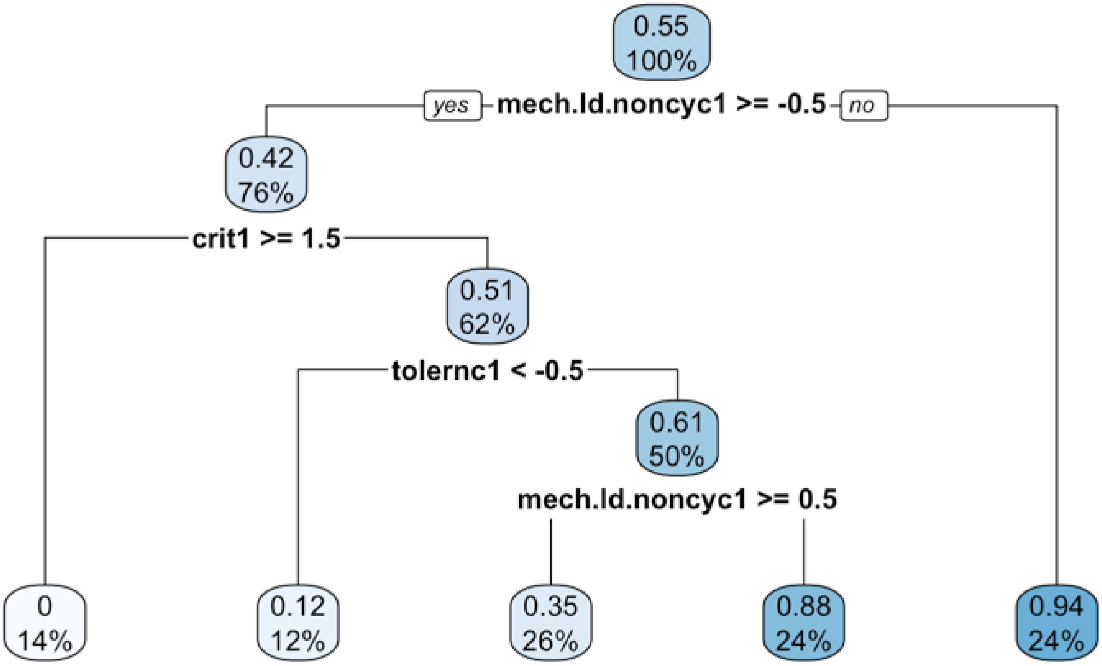


Figure 6.7: Decision tree for an expert using three attributes (with one repetition). Percentages indicate what percent of the overall choices are in a given branch of the tree while decimals represent the expected value of a given branch (where 1 represents choosing the part on the left and 0 choosing the part on the right in a pairwise comparison)

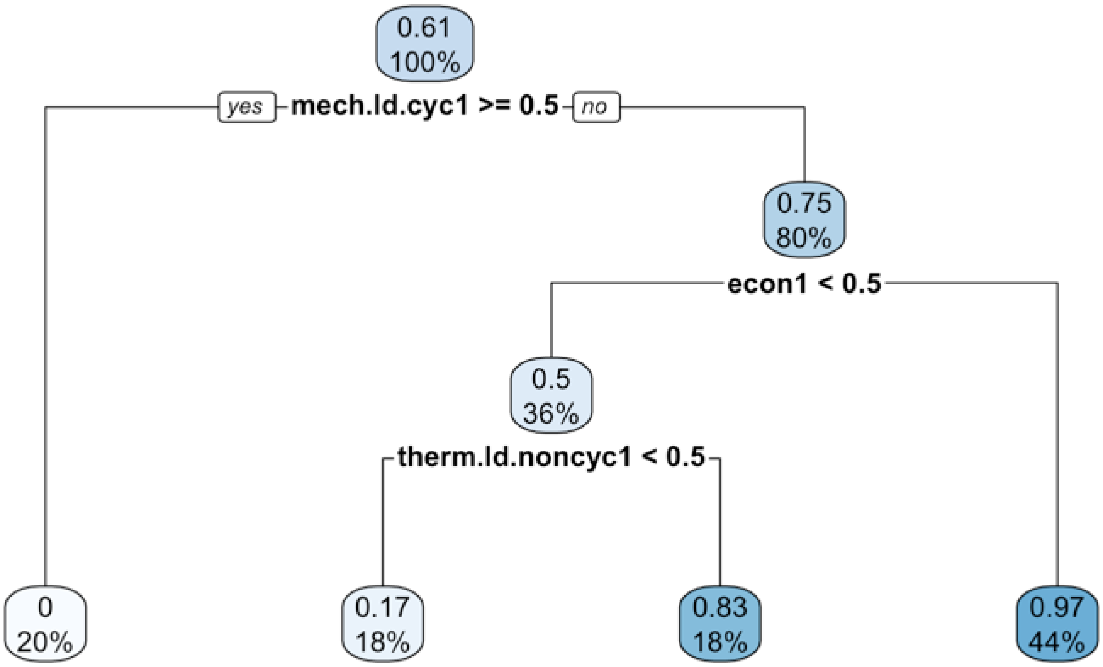


Figure 6.8: Decision tree for an expert using three attributes. Percentages indicate what percent of the overall choices are in a given branch of the tree while decimals represent the expected value of a given branch (where 1 represents choosing the part on the left and 0 choosing the part on the righ in a pairwise comparison)

## 6.2 Part descriptions

For a complete set of stimulus images, contact the corresponding author.

1. Engine Bracket (EB) – Mounting point to manipulate the engine in construction and repair
2. Fuel Swirler (FS) – Injection head mixes fuel and air for more complete combustion
3. Annular Combustor (AC) – Integrated ring of fuel injectors used as combustion point in most jet engines
4. Combustion Chamber Lining (CCL) – Casing for all combustion operations
5. Monocrystalline Turbine Blade (MTB) – Single crystal structured turbine blade forged to withstand extreme temperatures
6. Fan Disk (FD) – Structure to which blades are mounted for directing air flow in an engine
7. Thrust Reverser Blocker Door (TRBD) – Engine component at the rear of the engine used for engine braking
8. Turbofan Gearbox (TG) – Complex jet engine gear system which took over 30 years to design
9. Heat Exchanger (HE) – Chamber to transfer heat between fuel and oil
10. Vanes (V) – Stationary part to guide airflow in engine
11. Inlet Cone (IC) – Cone at the front of the engine to improve air flow into the engine chamber
12. Compressor Blisk (CB) – Portmanteau of blade and disk used for compressing air in the engine for increased combustion efficiency

## 6.3 Explicit questions

Listed below are all attribute questions asked in the second section of the survey.

1. Criticality: If this part fails, what is the likelihood it would cause catastrophic consequences in terms of damage to the aircraft or loss of life?

2. Overhangs: To what extent does this part contain sections that extend outward over the section below?
3. Tolerances: To what extent must this part's dimensions be precise?
4. Thermal Load - Cyclic: To what extent will this part be subjected to successive cycles of heating and cooling?
5. Thermal Load - Non-cyclic: To what extent will this part be required to operate at high temperatures?
6. Mechanical Load - Cyclic: To what extent will this part be exposed to mechanical loading and unloading, or to regular load reversals, while in operation?
7. Mechanical Load - Non-cyclic: To what extent will this part bear large mechanical loads?
8. Economics: To what extent can a business case be made to produce this part using additive manufacturing?

## 6.4 MDS fit

Permutation tests are one method of evaluating the fit of the MDS solution. By permuting the values of the dissimilarity matrix and computing an MDS fit on the permuted data we can see on average how often the stress from a random matrix would be as low as the one from the original fit from Figure 6.3. The stress value found for our data was 0.056 (Figure 6.2), falling well outside the distribution of random stress under permutation. A Shepard plot is another method of evaluating the fit of an MDS solution (Figure 6.4). Here we can see that the model fits well when parts are very similar or very different (0 and 0.5), but variation exists where the dissimilarity measure is less clear. All methods for MDS were implemented using the SMACOF package in R (de Leeuw & Mair, 2009).

## 6.5 Non-linear decision rule analysis

Due to 11 of our experts exhibiting quasi-complete separation in their survey responses, we were unable to model them using a stochastic discrete choice model. As the literature has shown that individuals often engage in tree-like decision practices (R. D. Luce, 1956; Tversky, 1972; Tversky & Sattath, 1979; Batley & Daly, 2006), we implemented a classification tree analysis using the rpart library in R (Therneau, Atkinson, & Ripley, 2015) to model their choice rules. Three examples of such choice rules can be seen in Figure 6.6, Figure 6.7, and Figure 6.8.

## 6.6 Sensitivity to excluded experts

We excluded three groups of respondents from the analysis of our results: experts who did not complete either portion of the survey (n=6), experts who only completed the first half of the survey (n=4), and experts who completed the survey, but failed either of the attention/expertise check questions (n=2). The first group cannot be included in a sensitivity analysis as none of the methods applied to the full group can be implemented without at least one half of the survey complete. Including the 6 experts who completed the first half of the survey at least we find that the average preference order changes by only a small degree. The thrust reverse blocker door and vanes change places, from fourth and third to third and fourth respectively. For the excluded expert group, every expert has a larger MFES than the average MFES from the original sample. Similarly, every expert in the excluded group has a larger HD than the average of the original sample. As may be expected, those who indicate some element of not being expert are more internally inconsistent and distant from the average response. Interestingly, this does not have a major impact on the average preference order. This result is supported by the simulation tests of HD and MFES both stabilize with increasing number of experts.

## 6.7 Coordinates in MDS space

Coordinates listed are (Economics, Tolerances, Thermal Loading – Cyclic). Parts are listed in the order returned by the discrete choice model.

1. Heat Exchanger (3.09, 3.33, 3.57)

2. Fuel Swirler (3.69, 3.57, 2.88)

3. Thrust Reverse Blocker Door (3.51, 3.16, 3.46)

4. Engine Bracket (3.29, 3.45,2.52)

5. Vanes (3.09, 3.02, 3.13)

6. Inlet Cone (2.85, 3.02, 3.73)

7. Annular Combustor (3.02, 3.51, 3.89)

8. Combustion Chamber Lining (2.58, 3.92, 3.43)

9. Monocrystalline Turbine Blade (2.88, 4.57, 3.81)

10. Fan Disk (3.04, 4.11, 3.88)

11. Compressor Blisk (2.76, 4.04, 3.94)

12. Turbofan Gearbox (2.83, 4.67, 3.54)

# References

Abdollahi, S., Davis, A., Miller, J. H., & Feinberg, A. W. (2018). Expert-guided optimization for 3d printing of soft and liquid materials. *PloS one*, *13*(4), e0194890. 5, 46

Ackerloff, G. (1970). The market for lemons: Quality uncertainty and the market mechanism. *Quarterly journal of economics*, *84*(3), 488–500. 2, 4, 25, 26

Ailon, N., Charikar, M., & Newman, A. (2008). Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, *55*(5), 1–27. 37, 40

Ali, I., Cook, W. D., & Kress, M. (1986). On the minimum violations ranking of a tournament. *Management Science*, *32*(6), 660–672. 33

Alon, N. (2006). Ranking tournaments. *SIAM Journal on Discrete Mathematics*, *20*(1), 137–142. 38

An, H. J., & Ahn, S.-J. (2016). Emerging technologies—beyond the chasm: Assessing technological forecasting and its implication for innovation management in korea. *Technological Forecasting and Social Change*, *102*, 132–142. 4

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, *3*(1), 20. 32

Autonomio talos [computer software]. (2019). Retrieved from `http://github.com/autonomio/talos.` 57

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*(4), 387–415. 33

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01 15

Batley, R., & Daly, A. (2006). On the equivalence between elimination-by-aspects and generalised extreme value models of choice behaviour. *Journal of Mathematical Psychology*, *50*(5), 456–467.

5, 82

Behler, J. (2011). Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Physical Chemistry Chemical Physics*, *13*(40), 17930–17955. 52

Behler, J., & Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, *98*(14), 146401. 52

Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological review*, *120*(3), 522. 49

Birnbaum, M. H. (2010). Testing lexicographic semiorders as models of decision making: Priority dominance, integration, interaction, and transitivity. *Journal of Mathematical Psychology*, *54*(4), 363–386. 50

Block, H., & Marschak, J. (1960). *Random orderings and stochastic theories of responses.* Contributions to probability and statistics, Olkin et al.(Eds.), Stanford . . . . 50

Bohn, R. E. (2005). From art to science in manufacturing: The evolution of technological knowledge. *Foundations and Trends(R) in Technology, Information and Operations Management*, *1*(2), 1-82. 2, 4

Bonnín Roca, J., Vaishnav, P., Morgan, M., Mendonça, J., & Fuchs, E. (2017). When risks cannot be seen: Regulating uncertainty in emerging technologies. *Research Policy*, *46*(7), 1215-1233. 4, 7, 25

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media. 14

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological review*, *113*(2), 409. 29

Bräuning, M., & Hüllermeier, E. (2017). Learning conditional lexicographic preference trees. In *Selected papers of the 3rd german-polish symposium on data analysis and applications* (p. 41). 50

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., . . . Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(04), 669–688. 52, 53

Broomell, S. B., & Bhatia, S. (2014). Parameter recovery for decision modeling using choice data. *Decision*, *1*(4), 252. 38, 51

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020).

*Language models are few-shot learners.* 51

Bureau, U. C. (2017). Exhibit 7, u.s. exports of goods by end-use category and commodity. 7

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432. 5, 26, 49

Cavagnaro, D. R., Gonzalez, R., Myung, J. I., & Pitt, M. A. (2013). Optimal decision stimuli for risky choice experiments: An adaptive approach. *Management science*, *59*(2), 358–375. 49

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural computation*, *22*(4), 887–905. 49

Chollet, F., et al. (2015). Keras. Retrieved from https://github.com/fchollet/keras 57

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05)* (Vol. 1, pp. 539–546). 52

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering  System Safety*, *163*, 109 - 120. doi: https://doi.org/10.1016/j.ress.2017.02.003 30, 40

Cooke, R., et al. (1991). *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand. 40

Cooke, R. M. (1988). Uncertainty in risk assessment: A probabilist's manifesto. *Reliability Engineering & System Safety*, *23*(4), 277–283. 30, 46

Coombs, C. H. (1964). A theory of data. 53, 55

Council, N. R. (2004). *Accelerating technology transition: Bridging the valley of death for materials and processes in defense systems*. Washington, DC: The National Academies Press. doi: 10.17226/11108 4

C Poulton, E., & R Freeman, P. (1966, 08). Unwanted asymmetrical transfer effects with balanced experimental designs. , *66*, 1-8. 10

Crane, D. (1969, 06). Social structure in a group of scientists: A test of the "invisible college" hypothesis. , *34*, 335. 2, 4, 25

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In *Simple heuristics that make us smart* (pp. 97–118). Oxford University Press. 32

Davis-Stober, C. P. (2012). A lexicographic semiorder polytope and probabilistic representations of choice. *Journal of Mathematical Psychology*, *56*(2), 86–94. 50

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 571–582. 5, 41

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106. 5, 25

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. doi: 10.1126/science.2648573 5, 38

de Condorcet, J. A. N., & Diannyére, A. (1797). *Esquisse d'un tableau historique des progrès de l'esprit humain*. 29

De La Maza, C., Davis, A., Gonzalez, C., & Azevedo, I. (2018). A graph-based model to discover preference structure from choice data. In *40th annual meeting of the cognitive science society (cogsci 2018)* (pp. 25–28). 49, 67, 69

de la Vega, W. F. (1983). On the maximum cardinality of a consistent set of arcs in a random tournament. *Journal of Combinatorial Theory, Series B*, *35*(3), 328–332. 38, 40

de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: Smacof in r. *Journal of Statistical Software*, *31*(3), 1–30. 14, 81

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Viad (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 461–493). American Psychological Association. 6, 30

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, *124*(4), 369. 49

Evers, N., Cunningham, J., & Hoholm, T. (1988). Bringing innovation to the marketplace. 2

Fischhoff, B. (1991). Value elicitation: Is there anything in there? *American psychologist*, *46*(8), 835.

49, 69

Fischhoff, B. (2005). Cognitive processes in stated preference methods. *Handbook of environmental economics*, *2*, 937–968. 49

Fischhoff, B. (2015). The realities of risk-cost-benefit analysis. *Science*, *350*(6260). 48

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232. 60

Funk, P., Davis, A., Vaishnav, P., Dewitt, B., & Fuchs, E. (2020). Individual inconsistency and aggregate rationality: Overcoming inconsistencies in expert judgment at the technical frontier. *Technological Forecasting and Social Change*, *155*, 119984. 31, 47, 49, 63, 67, 68

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*. 51

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, *105*(3), 387. 5

Geifman, Y., & El-Yaniv, R. (2017). Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*. 66

Genest, C., Zidek, J. V., et al. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, *1*(1), 114–135. 40

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, *103*(4), 650. 46

Goddard, S. T. (1983). Ranking in tournaments and group decisionmaking. *Management Science*, *29*(12), 1384-1392. doi: 10.1287/mnsc.29.12.1384 13

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1) (No. 2). MIT press Cambridge. 52

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, *12*(1), 19. 5

Halal, W. E., Kull, M. D., & Leffmann, A. (1998). The george washington university forecast of emerging technologies: a continuous assessment of the technology revolution. *Technological Forecasting and Social Change*, *59*(1), 89–110. 28

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Labs Technical Journal*, *29*(2), 147–160. 16

Hayek, F. A. (1945). The use of knowledge in society. *The American economic review*, *35*(4), 519–530. 2, 4, 25, 26

Ho, M. P., Gonzalez, J. M., Lerner, H. P., Neuland, C. Y., Whang, J. M., McMurry-Heath, M., . . . Irony, T. (2015). Incorporating patient-preference evidence into regulatory decision making. *Surgical endoscopy*, *29*(10), 2984–2993. 48, 67

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282). 60

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131. 5

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, *4*(2), 251–257. 51

Huang, R., Riddle, M., Graziano, D., Warren, J., Das, S., Nimbalkar, S., . . . Masanet, E. (2016). Energy and emissions saving potential of additive manufacturing: the case of lightweight aircraft components. *Journal of Cleaner Production*, *135*, 1559–1570. 24

Hurley, J. R., & Cattell, R. B. (1962). The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Systems Research and Behavioral Science*, *7*(2), 258–262. 14

ISO. (2010). Iso 286-1:2010(en) geometrical product specifications (gps) — iso code system for tolerances on linear sizes — part 1: Basis of tolerances, deviations and fits. 23

Jiang, R., Kleer, R., & Piller, F. T. (2017). Predicting the future of additive manufacturing: A delphi study on economic and societal implications of 3d printing for 2030. *Technological Forecasting and Social Change*, *117*, 84–97. 28

Judd, J. S. (1990). *Neural network design and the complexity of learning*. MIT press. 51

Kadane, J. B., & Fischhoff, B. (2013). A cautionary note on global recalibration. *Judgment and Decision Making*, *8*(1), 25. 5

Kamae, T. (1967). Notes on a minimum feedback arc set. *IEEE Transactions on Circuit Theory*, *14*(1), 78–79. 16

Keeney, R. L., Raiffa, H., et al. (1993). *Decisions with multiple objectives: preferences and value trade-offs*.

Cambridge university press. 47

Keeney, R. L., & von Winterfeldt, D. (1989). On the uses of expert judgment on complex technical problems. *IEEE Transactions on Engineering Management*, *36*(2), 83-86. 47

Kim, W., Pitt, M., Lu, Z., & Myung, J. (2017). Planning beyond the next trial in adaptive experiments: A dynamic programming approach. *Cognitive Science*, *41*(8), 2234–2252. 49

Kruskal, J. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129. doi: https://doi.org/10.1007/BF02289694 13

Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., . . . others (2019). A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. *Radiology*, *291*(3), 781–791. 41

Laureijs, R. E., Roca, J. B., Narra, S. P., Montgomery, C., Beuth, J. L., & Fuchs, E. R. (2017). Metal additive manufacturing: Cost competitive beyond low volumes. *Journal of Manufacturing Science and Engineering*, *139*(8), 081010. 23

Lee, M., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict nfl games.

30

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*(2), 57. 23

Luce, R., & Suppes, P. (1965). Utility, preference and subjective probability. *Handbook of Mathematical Psychology*, *3*, 249–410. 50

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica, Journal of the Econometric Society*, 178–191. 5, 82

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, *111*(30), 10984–10989. doi: 10.1073/pnas.1406138111 46

Marschak, J. (1959). Binary Choice Constraints on Random Utility Indicators.

5, 9, 25, 28, 50

McComb, C., Cagan, J., & Kotovsky, K. (2017). Optimizing design teams based on problem properties: computational team simulations and an applied empirical test. *Journal of*

*Mechanical Design*, *139*(4), 041101. 6, 30

McFadden, D. (1999). Computing willingness-to-pay for transport improvements. *Trade, theory and econometrics: essays in honour of John S. Chipman. Routledge, London*. 50

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior. 5, 14, 50

McNichols, D. (2008). *Tacit knowledge: An examination of intergenerational knowledge transfer within an aerospace engineering community*. University of Phoenix. 2, 7

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . others (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, *25*(5), 1106–1115. 5, 41

Moon, J. W. (2015). *Topics on tournaments in graph theory*. Courier Dover Publications. 33, 43, 45

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, *111*(20), 7176–7184. doi: 10.1073/pnas.1319946111 5, 46

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), 126–132. 46

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. 60

Polanyi, M. (1966). *The tacit dimension*. Garden City, NY: Doubleday. 4, 25

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532. 30, 40, 46

Price, D. J. D. S. (1963). *Little science, big science*. 2, 4, 25

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., . . . others (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*(1), 18. 38

Rao, S. J., Wang, Y., & Cottrell, G. W. (2016). A deep siamese neural network learns the human-perceived similarity structure of facial expressions without explicit categories. In *Cogsci*. 52

Regenwetter, M., & Davis-Stober, C. (2008, 01). There are many models of transitive preference: A tutorial review and current perspective. , 99-124. 9

Robinson, D. K., Lagnau, A., & Boon, W. P. (2018). Innovation pathways in additive manufacturing: Methods for tracing emerging and branching paths from rapid prototyping to alternative applications. *Technological Forecasting and Social Change*. 6

Roca, J. B., Vaishnav, P., Fuchs, E. R., & Morgan, M. G. (2016). Policy needed for additive manufacturing. *Nature materials*, *15*(8), 815. 7

Rogers, E. (1962). *Diffusion of innovations*. Free Press of Glencoe. 4

Russell, J. D., & Fielding, J. C. (2014). America makes: The national additive manufacturing innovation institute (namii) status report and future opportunities (postprint). 4, 8

Sawyer, J. (1966, 10). Measurement and prediction clinical and statistical. , *66*, 178-200. 5

Sergi, B., Azevedo, I., Xia, T., Davis, A., & Xu, J. (2019). Support for emissions reductions based on immediate and long-term pollution exposure in china. *Ecological Economics*, *158*, 26–33. 48, 49, 67, 69

Sergi, B., Davis, A., & Azevedo, I. (2018). The effect of providing climate and health information on support for alternative electricity portfolios. *Environmental Research Letters*, *13*(2), 024026. 48, 49, 61, 67, 68, 69

Shadish, W., Cook, T., & Campbell, D. (2002, 01). Experimental and quasi-experimental designs for generalized causal inference. 9

Sinuff, T., Adhikari, N. K., Cook, D. J., Schünemann, H. J., Griffith, L. E., Rocker, G., & Walter, S. D. (2006). Mortality predictions in the intensive care unit: comparing physicians with scoring systems. *Critical care medicine*, *34*(3), 878–885. 5

Spencer, J. (1980). Optimally ranking unrankable tournaments. *Periodica Mathematica Hungarica*, *11*(2), 131–144. 38

Stokes, D. E. (2011). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press. 6

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the ieee international conference on computer vision* (pp. 843–852). 51

Therneau, T., Atkinson, B., & Ripley, B. (2015). *rpart: Recursive partitioning and regression trees. r package version 4.1–10.* 82

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273. 5

Torgerson, W. (1958). *Theory and methods of scaling*. Wiley. 13

Torrano-Gimenez, C., Nguyen, H. T., Alvarez, G., & Franke, K. (2015). Combining expert knowledge with automatic feature extraction for reliable web attack detection. *Security and Communication Networks*, *8*(16), 2750–2767. 5

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press. 14

Trueblood, J., Brown, S., & Heathcote, A. (2013). The multi-attribute linear ballistic accumulator model of decision-making. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35). 50

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, *121*(2), 179. 49, 50, 53

Tsuchiya, A., Ikeda, S., Ikegami, N., Nishimura, S., Sakai, I., Fukuda, T., . . . Tamura, M. (2002). Estimating an eq-5d population value set: the case of japan. *Health economics*, *11*(4), 341–353. 46

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Tech. Rep.). EDUCATIONAL TESTING SERVICE PRINCETON NJ. 14

Tversky, A. (1969). Intransitivity of preferences. *Psychological review*, *76*(1), 31. 4, 28, 32, 46, 50, 53

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological review*, *79*(4), 281. 5, 82

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-31. 6, 30

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. 5, 26

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of business*, S251–S278. 5, 26

Tversky, A., & Sattath, S. (1979). Preference trees. *Psychological Review*, *86*(6), 542. 5, 82

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management science*, *39*(10), 1179–1189. 5, 26

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems* (pp. 3630–3638). 51

WA, K., FE, H., J, L., & et al. (1995). The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, *122*(3), 191-203. doi: 10.7326/0003-4819-122-3-199502010-00007 5

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 9. 51

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67–82. 51